Workshop Proceedings
Conference Committee

Program Chairs

Alessandro Micarelli, Roma Tre University, Italy
John Stamper, Carnegie Mellon University, USA

Organization Chair

Kitty Panoumpia, Neoanalysis, Greece

Workshops and Tutorial Chairs

Stephen E. Fanesali, Carnegie Learning, Pittsburgh, USA
Tsukasa Hirashima, Hiroshima University, Japan

Young Researcher Track Chairs

Darina Dicheva, Winston-Salem State University, USA
Toby Dragon, Ithaca College, USA

Award Chair

Claude Frasson, University of Montreal, Canada
The Conference is held under the auspices of the
Institute of Intelligent Systems

Sponsors

Springer

Best Paper Award

National Science Foundation

Where Discoveries Begin

Funding the YRT

Conference Organizers
<table>
<thead>
<tr>
<th>Time</th>
<th>ROOM 1</th>
<th>ROOM 2</th>
<th>ROOM 3</th>
<th>ROOM 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30</td>
<td>Registration</td>
<td>Registration</td>
<td>Registration</td>
<td>Registration</td>
</tr>
</tbody>
</table>
| 09:00 | W1 & W4: First International Workshop on Supporting Dynamic Cognitive Affective and Metacognitive Processes (SD-CAM) & 2nd International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2016)  
Jason M. Harley, Claude Frasson and Benedict du Boulay  
https://sites.google.com/site/sdcamworkshop/  
https://sites.google.com/site/amadl2016/ | W5: Building ITS Bridges Across Frontiers  
Stefan Trausan-Matu, Stefano Cerri and Mihai Dascalu  
Jennifer Olsen, Erin Walker, Roberto Martínuez-Maldonado, Ilya Golkin and Jihie Kim  
https://sites.google.com/site/ts2016islg/ | T1: Educational Data Analysis using LearnSphere  
Ran Liu, Michael Eagle, Philip Pavlik, John Stamper  
http://pskdata.shop.wvu.edu/ITS 2016/ |
| 10:30 | Coffee Break                                                            | Coffee Break                                                            | Coffee Break                                                            | Coffee Break                                                            |
| 11:00 | W1 & W4: First International Workshop on Supporting Dynamic Cognitive Affective and Metacognitive Processes (SD-CAM) & 2nd International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2016)  
Jason M. Harley, Claude Frasson and Benedict du Boulay  
https://sites.google.com/site/sdcamworkshop/  
https://sites.google.com/site/amadl2016/ | W5: Building ITS Bridges Across Frontiers  
Stefan Trausan-Matu, Stefano Cerri and Mihai Dascalu  
Jennifer Olsen, Erin Walker, Roberto Martínuez-Maldonado, Ilya Golkin and Jihie Kim  
https://sites.google.com/site/ts2016islg/ | T1: Educational Data Analysis using LearnSphere  
Ran Liu, Michael Eagle, Philip Pavlik, John Stamper  
http://pskdata.shop.wvu.edu/ITS 2016/ |
| 12:30 | Lunch                                                                  | Lunch                                                                  | Lunch                                                                  | Lunch                                                                  |
Vania Dimitrova, Lydia Lau, Amali Weerasinghe and Antonija Mitrovic  
Andrew Koster, Tiago Thompson Primo, Rosa Maria Vicari, Takao Terano and Fernando Koch  
http://www.digedu-workshop.org/ | T1: Educational Data Analysis using LearnSphere  
Ran Liu, Michael Eagle, Philip Pavlik, John Stamper  
http://pskdata.shop.wvu.edu/ITS 2016/ |
| 15:00 | W3: First International Workshop on Intelligent Mentoring Systems (IMS 2016)  
Vania Dimitrova, Lydia Lau, Amali Weerasinghe and Antonija Mitrovic  
Andrew Koster, Tiago Thompson Primo, Rosa Maria Vicari, Takao Terano and Fernando Koch  
http://www.digedu-workshop.org/ | T1: Educational Data Analysis using LearnSphere  
Ran Liu, Michael Eagle, Philip Pavlik, John Stamper  
http://pskdata.shop.wvu.edu/ITS 2016/ |
| 17:00 | End of sessions                                                         | End of sessions                                                         | End of sessions                                                         | End of sessions                                                         |
Introduction

ITS 2016 Workshops will be held on **June 7, 2016 in Zagreb, Croatia, in Hotel (Tomislav Hall).**

All selected workshops are Half day (HD) workshops. They cover a wide range of scientific topics varying in content, aim and structure in the field of Intelligent Tutoring Systems. In the context of the workshops, researchers from around the world with shared interests are offered an opportunity to:

- engage in professional exchanges leading to collaborations, grants and publications
- participate in the workshops in an interactive setting and a relaxed atmosphere which stimulates the development of ideas and interests.

The workshops are distinguished in:

- “classical workshops,” during which early work is presented in an informal, “mini conference” style,
- “working meetings” highly interactive for a group of people, designed to take advantage of being co-located to work together collaboratively on a theme of shared interest.

**ITS 2016 Workshops are:**

- **W1 (HD):** First International Workshop on Supporting Dynamic Cognitive Affective and Metacognitive Processes (SD-CAM)
  
  *Jason M. Harley and Claude Frasson*

- **W2 (HD):** 2nd International Workshop on Social Computing in Digital Education (SocialEdu 2016)
  
  *Andrew Koster, Tiago Thompisen Primo, Rosa Maria Vicari, Takao Terano and Fernando Koch*

- **W3 (HD):** First International Workshop on Intelligent Mentoring Systems (IMS 2016)
  
  *Amali Weerasinghe, Vania Dimitrova, Lydia Lau, and Antonija Mitrovic*

- **W4 (HD):** 2nd International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2016)
  
  *Benedict du Boulay, Beate Grawemeyer, Manolis Mavrikis, Genaro Rebolledo-Mendez, Olga C. Santos*

- **W5 (HD):** Building ITS Bridges Across Frontiers
  
  *Stefan Trausan-Matu, Stefano Cerri and Mihai Dascalu*

- **W6 (HD):** 5th Workshop on Intelligent Support for Learning in Groups (ISLG 2016)
  
  *Jennifer Olsen, Erin Walker, Roberto Martinwz-Maldonado, Ilya Goldin and Jihie Kim*

In addition to the selected workshops, a Full Day Tutorial will be held in a parallel track in the same conference venue. It will explore the application of novel educational data mining workflows using LearnSphere, a new data sharing and analysis portal.

**ITS 2016 Tutorial is:**

- **T1 (FD):** Educational Data Analysis using LearnSphere
  
  *(Ran Liu, Michael Eagle, Philip Pavlik, John Stamper)*
Table of Contents

Workshop 1: First International Workshop on Supporting Dynamic Cognitive Affective and Metacognitive Processes (SD-CAM)

Mission Accomplished? Measuring Gamers’ Emotion and Cognitive Engagement During and After a Narrative Event
   Jason M. Harley, Mohamed S. Benlamine, Maher Chaouachi, Claude Frasson, and Aude Dufresne

Learners’ Performance Tracking Using Eye Gaze Data
   Asma Ben Khedher, Imène Jraidi and Claude Frasson

Workshop 2: 2nd International Workshop on Social Computing in Digital Education (SocialEdu 2016)

Full Papers

SCALA on the Internet of Things: an Exploratory Research
   Roceli P. Lima, Magda Bercht, Liliana Passerino, Rosa Vicari and Joao Carlos Gluz

Towards a Conceptual Framework for Bridging Player Roles and Learner Roles in Gamified Collaborative Learning Contexts
   Simone Borges, Vinicius Durelli, Ig Ibner Bittencourt, Riichiro Mizoguchi and Seiji Isotani

The Role of Agent-based Simulation in Education
   Andrew Koster, Fernando Koch, Nicolas Riccieri Assumpção and Tiago Primo

Using Semantic Web Technologies to Describe an Educational Domain
   Tiago Primo, Helder Coelho, Rosa Vicari and André Behr

Short Papers

Model of assessment using tests formed of questions with different theoretical and practical difficulty degrees based on a genetic algorithm
   Doru Anastasiu Popescu and Nicolae Bold

The Impact of Social Similarities and Event Detection on Ranking Resources in Collaborative E-learning
   Samia Beldjoudi, Hassina Serdi and Abdallah Benzine

Workshop 3: First International Workshop on Intelligent Mentoring Systems (IMS 2016)
Position Papers: Underpinning Theories, Models, Technologies
Some Challenges for AIED Systems in Taking on Long Term Mentoring
Benedict Du Boulay

Intelligent Mentoring Systems for Making Meaning from Work Experience
Vania Dimitrova, Alexandra Poulavassilis, Nicolas Van Laebeke, Tamsin Treasure-Jones, Paul 
Brna and Miriam Zakas

Personalised and Adaptive Mentoring in Medical Education – the myPAL project
Nicolas Van Laebeke, Jane Kirby and Trudie Roberts

An Active Learning Model Employing Flipped Learning and Gamification Strategies
Darina Dicheva and Christo Dichev

Experimental Studies: Intelligent Mentoring Features

Dynamics of Trust in Group Peer Mentorship
Oluwabunmi Adewoyin, Fred Phillips and Julita Vassileva

Group Learning, Student Clustering and Peer Mentorship
Dorina Anastasiu Popescu, Ovidiu Domsa and Nicolae Bold

Usability of an Active Video Watching System for Soft Skills Training
Lydia Lau, Tanja Mitrovic, Amali Weerasinghe and Vania Dimitrova

Workshop 4: 2nd International Workshop on Affect, Meta-Affect, Data and 
Learning (AMADL 2016)

On the Feasibility of Providing Real-Time Adaptive Support for Motivation and Emotion in Intelligent Tutoring Systems
Seth A. Martin, Joseph F. Grafsgaard, Nicholas V. Mudrick, Michelle Taub, Roger Azevedo

Human Expert Labeling Process (HELP): Towards a Reliable Higher-Order User State Labeling by Human Experts
Sinem Aslan, Sinem Emine Mete, Eda Okur, Ece Oktay, Nese Alyaz, Utku Genc, David Stanhill, 
Asli Arslan Esme

Modeling Visual Attention of Students Playing an Educational Game for Physics
Celesamae T. Vicente, Ma. Mercedes T. Rodrigo, Hong-Fa Ho, Guan-An Chen

The coupling between metacognition and emotions during STEM learning with advanced learning technologies: A critical analysis and implications for future research
Roger Azevedo, Nicholas V. Mudrick, Michelle Taub, Seth A. Martin, Franz Wortha, and Garrett 
C. Millar

Workshop 5: Building ITS Bridges Across Frontiers

Choosing a Student Model for a Real World Application
Jiří Řihák and Radek Pelánek
Experiments with Students using an Intelligent Environment for Java
Maria Lucia Barron Estrada, Ramón Zatarain, Raúl Oramas Bustillos and Giner Alor Hernández

Towards the evaluation of electronic collaboration
Raouadha Chebil, Wided Lejouad Chaari and Stefano A. Cerri

Reusing assessments tests by generating arborescent test groups using a genetic algorithm
Nicolea Bold and Ovidiu Domsa
Workshop 1

First International Workshop on Supporting Dynamic Cognitive Affective and Metacognitive Processes

(SD-CAM)

https://sites.google.com/site/sdcamworkshop/
Supporting Dynamic Cognitive, Affective, and Metacognitive (SD-CAM) Processes

Jason M. Harley¹,² and Claude Frasson³

Université de Montréal, Computer Science and Operations Research, Montréal, QC, Canada ¹
University of Alberta, Educational Psychology, AB, Canada ²
j.harley1@ualberta.ca; frasson@iro.umontreal.ca

Preface

It’s time to focus on time. Specifically, the theoretical, methodological, and analytical considerations involved in accurately measuring and effectively adapting to learners’ evolving cognitive, affective, and metacognitive (CAM) processes over time. This workshop was organized to foster interdisciplinary discussion on these critical psychological processes for learning that have broad applications to academic and applied research (e.g., medical simulation, video game, aerospace, etc.).

Dynamic CAM data has not yet been harnessed and requires ongoing innovation through collaborative problem solving between research groups and disciplines with different skillsets in order to resolve a myriad of open questions [1-4]: Are our existing theoretical assumptions sufficient to guide research questions? What does agreement or disagreement between methods for a common CAM process mean? What limitations currently exist in measuring different CAM processes over time and are their ways to circumvent them? What sampling frequencies are most appropriate for measuring CAM processes over time? What happens to CAM processes over time? Are there common trajectories? Is there a common weave to CAM trajectories? And if so, for whom, and under what conditions?

Members of the ITS community have made great strides in advancing theoretical, methodological, and analytical solutions, as well as raising attention to the need for more solutions to be sought after [1-19]. As such, ITS and the SD-CAM workshop represented an ideal scientific community to focus and deliberate on the topic of time as it relates to CAM processes and explore the below topics in detail.

1.1 Topics of interest

- CAM transitions / traces / trajectories
- Relationship of context and stimuli / object foci to different temporal CAM patterns
- Temporal characteristics of CAM data
- Duration of CAM processes
- Sensors (EDA, EEG, Eye-tracking)
- Behavioural Data (posture, facial expressions)
• Latency of different methods

1.2 Workshop Goals
• Share international, interdisciplinary expertise on this complex topic.
• Foster a holistic discussion of temporal CAM processes that unites theoretical, methodological, and analytical considerations.
• Serve as a community resource for researchers, both established and emerging, to work collaboratively toward a sophisticated understanding of CAM.

2 Program committee (alphabetically listed)

2.1 SD-CAM Workshop Chairs
• Jason M. Harley (University of Alberta, Canada), Chair
• Claude Frasson (Université de Montréal, Canada), Chair

2.2 P.C. Committee
We wish to thank the program committee for their support and reviews which made the first SD-CAM Workshop possible.
• Ivon Arroyo (Worcester Polytechnic Institute, USA)
• Roger Azevedo (NC State, USA)
• Ryan Baker (Teachers College, Colombia University, USA)
• Gautam Biswas (Vanderbilt University, USA)
• Benedict du Boulay (University of Sussex, U.K.)
• François Bouche (Sorbonne University / CNRS, France)
• Cristina Conati (University of British Columbia, Canada)
• Maher Chaouachi (McGill University / CAE, Canada)
• Joseph Grafgaard (NC State, USA)
• Sazzad Hussain (University of Sydney / CSIRO, Australia)
• Eric Poitras (University of Utah, USA)
• Blair Lehman, (Educational Testing Services, USA)
• Suanne Lajoie (McGill University, Canada)
• James Lester (NC State, USA)
• Kasia Muldner (Carleton University, Canada)
• Reinhard Pekrun (University of Munich, Germany)
• Jonathan Rowe (NC State, USA)
References


Mission Accomplished? Measuring Gamers’ Emotion and Cognitive Engagement During and After a Narrative Event

Jason M. Harley¹², Mohamed S. Benlamine², Maher Chaouachi², Claude Frasson², and Aude Dufresne²

University of Alberta, Educational Psychology, AB, Canada ¹
Université de Montréal, Computer Science and Operations Research, Montréal, QC, Canada ²
jharley1@ualberta.ca

Abstract. This study examined the synchronization of multimodal affective and cognitive engagement data during and immediately following a narratively-significant event in a gaming session with Assassin’s Creed: Unity. Data was collected using an electroencephalography (EEG) headset, electrodermal activation (EDA) bracelet, and a self-report measure of emotions from 20 university students. Results revealed: (1) a lack of agreement between EDA and self-report data regarding gamers’ affective arousal; (2) small and marginally significant negative correlations between EDA arousal and cognitive engagement from EEG data; (3) no clear directional trend in changes in physiological data pre-to-post the narratively-significant event; and (4) that the average number of hours an individual spent playing games a week influenced their physiological states, especially their affective arousal. Cumulatively, results challenge theories and results that purport a relationship between (1) emotional expression components, and (2) between cognitive engagement and affective arousal, as well as provide preliminary evidence of physiological trends in users based on the quantity of hours they spend playing games.

Keywords: Emotions, affect, multimodal measurement, affect detection, games
Learners’ Performance Tracking Using Eye Gaze Data

Asma Ben Khedher, Imène Jraidì and Claude Frasson
Université de Montréal
Département d’informatique et de recherche opérationnelle
2920 Chemin de la Tour, Montréal, H3T-1H, Québec, Canada
{benheda, jraidiim, frasson@iro.umontreal.ca}

Abstract. Endowing computer-based learning environments with the faculty of tracking a learner’s eye gaze behavior could provide a deeper insight into the learner’s interaction experience and improve the communication and interaction methods. In this paper we analyze the relationship between the learners’ eye gaze patterns and their performance while interacting with Crystal Island, a narrative-centered learning game environment. Experimental data were gathered from 20 participants using an eye tracker while they were reading virtual books and answering multiple-choice quizzes. Results showed clear trends in the eye gaze behavior across the learners answer attempts and significant impacts on response time, accuracy and progress within the game. Our findings have implications for technology-enhanced learning environments seeking to continuously assess and adapt to the learner’s experience.

1 Introduction

One of the major interests in computer-based learning environments is to provide an adequate support to the learners to help them overcome their learning shortcomings [1, 2]. Therefore, it is important to evaluate the quality of the students’ interaction experience and measure their progressions. There is also a need to monitor the learners’ cognitive performance in order to figure out whether they have well acquired the new knowledge. One of the current approaches in assessing the learners’ experience is the use of non-intrusive sensors, as they provide valuable quantitative and unbiased information as compared to traditional methods such as questionnaires or self-reports [3, 4, 5, 6, 7].

Eye tracking is a promising technique to use in this context since it provides an objective monitoring of the learners visual activity [8]. Indeed, eye tracking devices have been widely used in human-computer interaction and experimented in different fields including visualization [9, 10], activity recognition [11, 12], and affect detection [13, 14]. Many researchers have also used eye tracking within educational settings [15, 16, 17, 18]. Kardan and Conati [18] used eye tracking to assess users learning and discriminate between two classes of learners: high and low learners. In [16], Lin investigated the effect of geometry problem’s difficulty on gaze data indicating that the more the problems were difficult, the more the fixations on the geometric figures were long-lasting.

© Springer-Verlag Berlin Heidelberg 2011
In this paper we propose to enhance a serious game environment (Crystal Island) with eye tracking metrics. We seek to investigate how a learner’s eye gaze behavior relates to task performance. An experimental study was conducted to record learners’ eye gaze data while reading virtual books and acquiring knowledge about microbiology. Results demonstrated identifiable trends across the learners answer attempts and showed significant relationships between the eye gaze features and learners’ responses and progress.

The organization of this paper is as follows: first we present some related works on eye gaze usage. Then, we describe our learning environment and the experimental protocol. After that, we detail and discuss the obtained results and the future work.

2 Related work

Several studies showed the effectiveness of using gaze data as an information source about the users’ cognitive performance, since eye gaze is closely related to human cognition and brain activity [19, 20, 21]. For instance, the main eye tracking metrics such as fixations (i.e. keeping the eye gaze on a particular point of the screen) and saccades (i.e. a rapid eye movement from a point of fixation to another) are indicators of the amount of mental effort being provided during the visual process. Fixation duration reflects processing time and difficulty: the longer a user is looking at a specific information, the more it was difficult for him to understand it [22]. Chen et al. [19] conducted an experiment where students’ eye gaze were recorded as they responded to physics concepts presented in two modality (text and picture). The results revealed that the re-reading time metric indicates comprehension difficulty. Steichen et al. [23] explored gaze data in information visualization to infer users’ cognitive abilities such as working memory. The study showed that eye gaze can reliably predict users’ tasks and cognitive abilities.

Within learning environments, tracking students’ eye gaze is gaining importance as a method to assess users’ learning performance and skills. Kordan and his colleagues [18] used eye tracking to discriminate between two classes of learners (high and low learners) with the aim to provide early interventions. In [24], the authors investigated users’ visual behavior and showed that students’ performance and individual cognitive abilities can be inferred from eye gaze data. In a reading activity, Martínez-Gómez and his colleagues [25] investigated whether eye gaze data as well as reading behavior are predictive of language skill and level of understanding. The results demonstrated that lower level students have larger average fixation duration and smaller saccade median length.

Toker et al. [26] and Lallé et al. [27] examined eye gaze information as a data source to predict learners’ skill acquisition during information visualization.

The present work seeks to add to this body of work by analyzing users’ eye gaze behavior within a serious game environment. We explore the relationship between different eye gaze patterns and learners’ performance.
3 CRYSTAL ISLAND

Crystal Island is a task oriented learning serious game in the domain of microbiology [28]. The events take place in a volcanic island where the members of a research team were contaminated by an unidentified infectious disease. Players are attempting to discover the cause of this illness. They are free to perform different actions in order to collect the maximum number of clues. They can interact with the game’s characters, run experiments in the laboratory and also read books to obtain further information about the diseases and the infections. Once all the relevant information have been gathered, the final step is to choose among candidate diagnoses (botulism, cholera, etc.) the one that matches the patients’ symptoms and analysis results, as well as the appropriate treatment.

During the game, learners encounter several books and journals that they should read to deepen their knowledge. Some of them contain mandatory quizzes that they have to answer to continue their discovery. In this study we focus on this part of the game, we aim to test if the participants assimilated the new concepts covered in the books and if they were able to answer the questions correctly.

4 Experimental study

In this experiment, we recorded participants’ gaze data as they were reading books and solving multiple-choice quizzes. A commercial eye tracker (Tobii Tx300) with a sampling rate of 300 Hz was used. Free head movements were allowed during the experiment. EEG signals were also recorded during the experiment for future analyses. We used the biometric Attention Tool platform of iMotions for eye gaze visualization and data synchrottization.

<table>
<thead>
<tr>
<th>Table 1, Statistics of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Respondents</td>
</tr>
<tr>
<td>Average age</td>
</tr>
<tr>
<td>Std. Dev</td>
</tr>
<tr>
<td>Data quality</td>
</tr>
</tbody>
</table>

4.1 Participants

Data were collected from twenty graduate students recruited from the computer science department. Participants were between 17 to 33 years old (M = 25.9, SD = 4.34). Table 1 shows the age distribution of the participants as well as the gaze data quality
based on the calibration results. Data from one participant were discarded, as he did not read any book.

### 4.2 Protocol

Upon arrival at the laboratory, participants were briefed about the experimental objectives and procedure and asked to sign a consent form. They were then outfitted with an EPOC headset and placed in front of the eye tracker. Each session started with a calibration step and ended with a post-test.

**Calibration.** Tobii eye tracker uses a standard 9-point calibration grid. The calibration process, which lasts for about 20 seconds, consists in evaluating the gaze points’ quality. All participants in the experiment passed the calibration phase successfully.

**Session.** After the setup process, participants started using Crystal Island. They could explore the island differently, and were free to read all the books or some of them. Thus, the number of read books varied from one learner to another. On average participants read 5.68 books (SD = 1.733). Considering the books with quizzes, the number of questions per quiz is also variable. Each quiz contains on average 3.46 questions (SD = 0.97). The player has 50 minutes of time to resolve the mystery and identify the cause of the disease.

**Post-test.** This test contains twelve multiple-choice questions related to different scientific concepts covered in the game. The goal is to test the participants’ level of acquired knowledge after their interaction with the game. For each question, there is one correct alternative: 1 point for a correct choice and 0 point for a wrong one.

**Table 2.** Eye gaze features

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean fixation duration (ms)</td>
<td>The mean of all the fixation points’ duration in a book,</td>
</tr>
<tr>
<td>Number of fixations</td>
<td>The number of fixations recorded when reading,</td>
</tr>
<tr>
<td>Number of revisits</td>
<td>The number of times the learner returned to re-fixate the book,</td>
</tr>
<tr>
<td>Total fixation duration (ms)</td>
<td>The sum of all the fixation durations for all the books,</td>
</tr>
</tbody>
</table>

### 4.3 Data analysis

Participants have three attempts per quiz; some of them succeeded at first attempt and others could not answer the questions even after three attempts (M = 1.564, SD = 0.40). For each participant we computed detailed and global performance features.
The detailed performance features included the accuracy of participants’ responses and the response time, and were computed per quiz and within each attempt. As global features, we computed the learner’s game progress and the post-test’s results.

Participants’ eye gaze was recorded during the whole session. However in this study we were only interested in analyzing the gaze data while reading the virtual books and answering the quizzes. Four features were used as described in Table 2. Three features were extracted from each book: mean fixation duration, number of fixations and number of revisits; this latter determines how many times the participant looked back at the book when answering the questions (i.e. re-read the text). The fourth feature: total fixation duration, was summed over all the read books.

5 Results and discussion

The experimental results are presented in the following subsections. The first part is concerned with the learners’ performance while responding to the quizzes in the virtual books. The second part deals with the learners’ post-test results and progress within the game.

5.1 Quiz performance

Our first objective was to assess the impact, if any, of the learners’ eye gaze behavior on their learning outcomes. More precisely, we wanted to analyze the relationships between the learners’ eye tracking features during the book reading activity and their performance while solving the multiple-choice questionnaires. For each learner’s first answer attempt, we looked at the mean fixation duration (MFD), the number of fixations and the number of revisits.

Response accuracy. Three distinct analyses of variances (ANOVA) were conducted in order to check whether there were significant differences in terms of MFD, number of fixation and number of revisits respectively, between the learners who succeeded (i.e. answered correctly to all the questions related to the virtual book in their first attempt: response accuracy = 100%) and the learners who failed (response accuracy < 100%).

The first ANOVA showed a statistically significant difference between the two groups, $F(1, 106) = 6.642, p < 0.05$, revealing a significant relationship between the learners’ MFD during the reading phase and their answers. We have found that the learners who completed the quiz in the first attempt had in average a significantly lower mean fixation duration while reading ($M = 277.83, SD = 48.09$) than the learners who failed ($M = 301.38, SD = 43.93$). This result goes in line with previous studies which suggest that fixation duration is related to processing difficulties. Indeed the more the learners looked blankly at the screen, the more they were likely to have trouble understanding the concepts seen in the visited books and thus responding to the questionnaires.

The two remaining ANOVA were not statistically significant, $F(1, 106) = 3.347, p = n.s.$ for the number of fixation, and $F(1, 106) = 1.138, p = n.s.$ for the number of
revisits. This suggests that there was no reliable relationship between fixation count and accuracy.

**Table 3. Bivariate correlational results of the first answer attempts**

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean fixation duration</td>
<td>0.219*</td>
<td>0.023</td>
</tr>
<tr>
<td>Number of fixations</td>
<td>0.811**</td>
<td>0.000</td>
</tr>
<tr>
<td>Number of revisits</td>
<td>0.725**</td>
<td>0.000</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

**Response time.** Three bivariate correlations were run between learners’ response time during their first attempt answering the quizzes and the three eye tracking features measured while reading, namely MFD, number of fixations and number of revisits respectively. The three Pearson’s correlation coefficients (r) were statistically significant revealing positive linear relationships among the targeted variables as shown in Table 3. In particular, the correlation between response time and MFD was low (r = 0.219). The highest correlation coefficient was found between response time and the number of fixations (r = 0.811) suggesting that the more the participants looked repeatedly at the books, the more they tended to spend time on the questionnaires.

The number of visits seems thus to have a higher impact as compared to the duration of fixations, which suggests that the reading style could also impact learners’ performance. This was confirmed by the correlation between response time and the number of revisits, which was also high (r = 0.725). The learners who frequently visited (or revisited) the virtual books took their time to respond, as compared to the other learners who read the books less repeatedly and hence might have understood the concepts faster (fewer fixations) and had less difficulty answering (lower response time).

These correlational analyses were replicated in the second and third question answer attempts. We found the same patterns as in the first learners’ attempts. Results are shown in Table 4. The three correlations were statistically significant and the highest Pearson’s correlation coefficients were found between response time and the number of fixations.

**Eye gaze features across the attempts.** Repeated measures ANOVA were performed in order to analyze how the eye gaze features (MFD, number of fixations, and number of revisits) varied across the three answer attempts. Results showed a significant change in the eye gaze behavior between the first and the second response attempts, and no significant change between the second and the third attempts (p – n.s.). As depicted in Figure 1, the three measures decreased considerably from the first to the
second answer attempts: F (1, 42) = 77.019, p < 0.001 for the mean fixation duration, F (1, 42) = 80.882, p < 0.001 for the number of fixations, and F (1, 42) = 44.890, p < 0.001 for the number of revisits.

At first glance, one might think that this result is counter intuitive, as learners are expected to provide at least as much effort in their second attempts, but in fact this is consistent with the design of the quiz. Indeed during the second (and also the third) attempts, the learners are asked to retake only the missed questions, so that they can move to the next step of the game. In the virtual books, the learners are thus likely to revise only the concepts related to those questions.

**Table 4.** Bivariate correlational results of the second and third answer attempts

<table>
<thead>
<tr>
<th></th>
<th>2nd attempt (N = 43)</th>
<th>3rd attempt (N = 19)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>p</td>
</tr>
<tr>
<td>Mean fixation duration</td>
<td>0.617**</td>
<td>0.000</td>
</tr>
<tr>
<td>Number of fixations</td>
<td>0.811**</td>
<td>0.000</td>
</tr>
<tr>
<td>Number of revisits</td>
<td>0.716**</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2 tailed).**

5.2 Game performance

In this part we evaluate the relationship between eye gaze fixation and learners’ game performance. In particular we analyze the impact of the total fixation duration on the overall progress within the game and outcomes in the post-test.

**Game progress.** As previously mentioned, total fixation duration (TFD) is the sum of all the fixations measured during the book exploration phases. A bivariate correlation between TFD and the learners’ overall progress rate was performed. A statistically significant Pearson’s correlation coefficient was found showing a moderate negative relationship between the two variables (r = -0.567, p < 0.05). This suggests that the longer the TFD was, the lower the learners advanced in the game. In other words, the more the learners spent time exploring the virtual books, the more they were likely to struggle discovering the enigma of the game. In fact, 82% of the participants could not achieve this goal.

**Post-test results.** A bivariate correlation was run between TFD and the post-test outcomes. Results showed that there was no significant relationship between the two variables (p = n.s.). Indeed the post-test included a variety of questions related to different concepts seen in the game. For instance, a learner could spend a lot of time reading a particular book, studying and assimilating its concepts, but these latter were not included in the post-test. On the other side, a learner may not linger in the books,
but perfectly responds to the questions since he had assimilated new knowledge during his interaction with the game.

![Bar chart showing mean fixation duration, number of fixations, and number of revisits across three attempts.](image)

**Fig. 1.** Estimated marginal means of the eye gaze features over the learners' attempts. Error bars denote standard error.

### 6 Conclusion

An experimental study was conducted in order to explore the students' visual behavior while interacting with a serious game environment. An eye tracking device was used as learners were reading virtual books and solving multiple-choice quizzes. The goal was to check whether eye gaze can give insights into a learner's progress and gaps within the game.

Results revealed clear trends showing that the eye gaze measures vary in accordance with the design of the activity across the learners' attempts. Significant correlations were found between eye gaze and learning performances. We have shown that longer fixation duration was related to poorer performance in terms of accuracy in the questionnaires. This was confirmed at the global level of the game. It was found that fixation duration has also a negative impact on the learners' progress rates within the game. This is important considering the fact that learning performance and outcomes are of primary interest in the design of learning technology. For instance, in our case, eye tracking can be used to trigger assistance to help students while answering the questions whenever long fixations are noticed.

Correlational analyses showed that the reading style (longer fixation vs. repeated fixations) can also impact the learners' responses time. In our future work, we aim to broaden this latter analysis. Further variables such as the saccades and the fixation rates will be investigated.
Acknowledgment

We acknowledge SSHRC (Social Science and Human Research Council) through the LEADS project and NSERC (National Science and Engineering Research Council) for funding this research. A special thanks to James Lester and Roger Azevedo from University of North Carolina for their collaboration.

References

Workshop 2

2nd International Workshop on Social Computing in Digital Education
(SocialEdu 2016)

http://www.digedu-workshops.org/
Preface

Andrew Koster¹, Tiago T. Primo¹, Rosa Vicari², Takao Terano³, Fernando Koch¹

¹SAMSUNG Research Institute Brazil, Campinas SP, Brazil
{andrew.k, tiago.t, fernando.koch}@samsung.com
²Federal University of Rio Grande do Sul, Porto Alegre RS, Brazil
rosa@inf.ufrgs.br
³Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
terano@dis.titech.ac.jp

Digital Education is a theme as old as the first computer. The use of computers for teaching and learning activities are constantly associated with improvement of learning performance measured through parameters like level of attention and knowledge acquisition. However, there is a longstanding dispute on what pedagogical activities, class contexts and format of classroom instrumentation play a better role in maximizing the benefits of applying technology in education.

With that in mind, the Second International Workshop on Social Computing in Digital Education promotes the discussion on how social computing plays a role in the context of education supported by technology. This is a growing field of research concerned with bringing interactive social intelligence to teaching and learning. The event challenges the participants to envisage new forms of automatic assessment, models to understand and quantify learning performance, and analysis around the impact of socio-behaviour and context information on knowledge acquisition and retention. We seek to answer how cutting-edge research into social computing is applied in digital education, and are stimulating to those concerned with bringing interactive social intelligence for teaching and learning in the classroom.

A basic premise is to think of multidisciplinary investigative approaches. The work address topics such as: content selection and recommendation; gamification in education; advanced learning analytics; computational intelligence and machine learning; agent-based modelling and simulation; knowledge management, learning environment; smart interaction.

The Program Committee selected six contributions for these proceedings representing a broad range of cutting-edge areas in social computing and artificial intelligence applied in diverse education scenarios.

We would like to thank all the volunteers who made the workshop possible by helping to organise and peer reviewing the submissions. In addition, we are grateful to the ITS conference for providing the opportunity for the second edition of this workshop.
Program Committee

Sergio Alvarez-Napagao
Ricardo Azambuja Silveira
Raimundo Barreto
Marta Rosecler Bez
Ig Ibert Bittencourt
Carlos Cardonha
Juan Colonna
Néstor Dario Duque Méndez
Christian Gutmann
Seiji Isotani
Pablo Noriega
Eduardo Oliveira
Hugo Paredes
Alexandre Ribeiro
Jordi Sabater-Mir
Elder Santos
Jaime Siechman
Joao Luis Silva
Jose Viterbo
Leandro Krug Wives
SCALA on the Internet of Things: an Exploratory Research

Roceli P. Lima\textsuperscript{1}, Magda Bereht\textsuperscript{1}, Liliana M. Passerino\textsuperscript{1}, Rosa M. Vicari\textsuperscript{1}, João C. Gluz\textsuperscript{2}

\textsuperscript{1}Interdisciplinary Center for Educational Technologies (CINTED) – UFRGS – Brazil
roceliflima@gmail.com, bercht@inf.ufrgs.br, liliana@cinted.ufrgs.br, rosa@inf.ufrgs.br

\textsuperscript{2}Post-Graduation Program in Applied Computer Science (PPCA) – UNISINOS – Brazil
jcgiluz@unisinos.br

Abstract. The purpose of this work is to introduce SCALA system, an Alternative Communication System for the Literacy of People with Autism, and show how to integrate SCALA on the Internet of Things (IoT). After introducing SCALA, the work presents a review on the state of art about the application of IoT technologies to Autism, carried out in five international scientific databases. Based on this study, it is proposed a model for integrating SCALA to IoT with a focus on questions about the literacy process for children with autism. SCALA currently operates on Web, Desktop and Tablets platforms. The proposed technological integration with IoT aims to support additional functionalities of SCALA, which will enable mediation with its subjects by means of physical objects.

Keywords: Internet of Things, Autism, Alternative Communication, Mediation

1 Introduction

From 2009 onwards, the SCALA system, an Alternative Communication System for the Literacy of People with Autism [15], is being developed by our research group. During this period SCALA has been adapted to operate on several digital platforms, starting from desktop and laptop computers, passing by tablets and, more recently, supporting a fully web interface [15]. The next challenge is to integrate this system to the Internet of Things.

The expression Internet of Things (IoT) was coined in 1999 at a lecture by researcher Kevin Ashton for P&G executives, focused on the idea of embedding electronic devices in physical objects, allowing this objects to established communication links to the Internet [3, 8]. A follow up on this idea, resulted in the creation of Radio-Frequency Identification (RFID) standard [19].

Another evolution taking place during the same time, occurred with the use of Information and Communication Technologies (ICT) with children diagnosed with Autism Spectrum Disorder (ASD). This disorder has two domains: social/communication deficits, as well as fixed interests and repetitive behaviors (see
DSM-5 [9]). Among the ICTs used with children with ASD, we highlight the use of robots [2, 17, 13], alternative communication systems on computers [15], on the web [6,7], on tablets [11], games, apps for mobile platforms, tables with tangible user interface, as well as virtual platforms with 2D and 3D animations [10, 12, 14].

To understand better this complex set of applications and technologies, which forms the context of the integration of SCALA with IoT, we made a systematic exploratory search over five international scientific databases, to find recent works about the integrated application of IoT technology to work with Autism cases. The results of this research will help us to define a clear road map on how to evolve SCALA to work on the IoT.

This article is divided into 5 sections. In Section 2, the SCALA System is introduced, highlighting technological evolutions and functionalities. Section 3 the methods and results of the exploratory search on the intersection of IoT and Autism. Section 4 a proposal for integrating SCALA on the IoT. Section 5 final comments about the results achieved until now and about the future steps of this work.

2 The SCALA System

SCALA is an alternative communication system for the literacy of people with autism with no oral skills or with severe motor handicap [15]. The system was initially focused on children aged from 3 to 5.

Autism is a disorder affecting neurological development and it must be present since childhood or early childhood. According to the American Psychiatric Association, DSM-5, autism included in the diagnosis category for neurodevelopment disorders, under the sub-category of Autism Spectrum Disorders, which includes autism, Asperger's syndrome, Childhood Disintegrative Disorder (CDD) and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) [9].

The SCALA desktop system [4] was designed as a tool to build and use communication boards, which can be used as alternative communication devices with children with Autism. SCALA communication boards support the use of communication resources such as pictographs, voice synthesizing, audio recording, subtitles and animation. Lately, SCALA was evolved to support web and tablet platforms [7]. New functionalities also were incorporated in the system [7]: the Visual Narratives module, the prototype for a free communicator device, and a scanning system for the communication board module. Throughout the development process, several studies were developed involving children between 3 and 5 years with autism, their families and teachers of the schools. SCALA was entirely designed and developed using the Context-Centered Design of Usage principles [15], which allowed us to identify points to be improved and re-engineered.

Current research on SCALA, besides the integration with IoT, includes, among other works, the development of a formal ontological support for the semantics of the alternative communication system, the integration and use of ARASAAC pictographs (http://catedu.es/arasaac/) into SCALA, and the support of a scanning mode in order to include individuals with physical disabilities.
The development of web version of SCALA system (http://scala.ufrgs.br/Scalaweb/) begins in 2011 with the proposal to consider a cross-platform solution especially designed to meet the demands of applications for tablet. All development was accompanied and marked with a study case with three 3-4 years old children with Autism over two years. This study was developed in three different contexts: school, family and lab at the university. The SCALA has two modules: Communication Board (Figure 2) and Visual Narratives (Figure 3) and was developed under GNU and Creative Commons licenses to ensure its open content.

The pictographs used in the system were mostly developed by ARASAAC (http://catedu.es/arasaac/) group, and are available under Creative Commons licenses. With the use of ARASAAC images and with the users’ own pictures, SCALA
currently has more than four thousand images, divided into the categories: People, Objects, Nature, Action, Food, Feelings, Qualities and My Pictures, where the user has the option inserting his own images in the system.

The modules differ in some ways that allow greater flexibility, depending on the objectives, strategies and degree of difficulty to be proposed for use. The Communication Board module has a user friendly interface that integrates in a single screen facilities to edit and use communication boards, allowing teachers, tutors or parents to create boards that can be immediately used to communicate with children. In this module there is static space on the screen where the user has the possibility of choosing a layout for construction of single boards (Figure 2). In Visual Narratives module spaces are free to insert multiple images and textual elements (Figure 3). The Communication Board module consists of an editor of communication boards and a viewer for these boards with built-in voice synthesizer. Communication boards are assistive technology resources to facilitate communication between people with deficits in communication and other participants from the signaling or choice of images previously organized in the boards.

In SCALA system communication boards start from a scalable basic layout, which allows every user to create their own boards simply and directly. Each pictograph in a board is editable, allowing to change its caption and even the sound of it from a context-sensitive menu. The boards can be stored privately or in public spaces.

![Figure 3. SCALA Web – Visual Narratives module interface.](image)

The Visual Narratives module (Figure 3) allows building stories with flexible preparing conditions. This module has diverse layouts that provide a variable degree of complexity. The screen has a blank space where it is possible to insert and edit images, which can be overlapped, increased or decreased in size, inverted or deleted. There is the possibility to define the color background and insert and edit conversation balloons with small dialogues. The story then can be saved for posterior use. It also can be played by the tablet's voice synthesizer which will read what was typed, otherwise the recorded sounds will be played.
3. Review of the State of Art on IoT and Autism

The review of the state of the art started with a systematic literature review based on the exploratory research method devised in [Gil, 1991], which aims to develop, clarify and change concepts and ideas, especially with the purpose to specify more accurately problems and hypotheses for subsequent studies.

The exploratory research method starts with the choice of the national and international scientific databases to be searched (see Section 3.1), follows to the definition of the descriptors to be used on the search of these databases (see Section 3.2) and to the execution of the search and the selection of works and articles (see Section 3.2). Detailed results about the most representative articles found about the application of IoT to Autism are presented in Section 3.3.

3.1 Scientific databases

For the execution of the systematic research, five international scientific databases were initially analyzed, two of which pertaining to the field of Exact and Earth Sciences, and the three others pertaining to Multi-Disciplinary fields, ranging from 1999 to 2015. The selected databases of Exact and Earth Sciences: Association for the Computing Machinery (ACM) and Institute of Electrical and Electronic Engineers (IEEE). The Multi-Disciplinary selected: Science Direct, Scopus and Springer.

3.2 Descriptors of the exploratory research

The exploration criteria defined for the research was to select articles from the databases that meet a query descriptor including the terms Internet of Things and Autism. As shown in Table 1, the bibliographical research found 13 articles based on this descriptor, 5 of which obtained from the IEEE database, 1 from the SCOPUS database and 7 from the SPRINGER database.

However, only three articles were considered for the detailed analysis, being considered the most representative of current works on the application of IoT technology to Autism. The contents of 2 articles were already included in other articles extracted from the databases. Other 8 articles did not present an integrated solution in terms of Internet of Things and Autism and were not considered.

<table>
<thead>
<tr>
<th>#</th>
<th>Source</th>
<th>&quot;Internet of things&quot; + &quot;autism&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ACM</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>IEEE</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>SCIENCE DIRECT</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>SCOPUS</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>SPRINGER</td>
<td>7</td>
</tr>
</tbody>
</table>
3.3. Results of the Exploratory Research

The only work to make use of a machine learning algorithm was [5], which presented a web information retrieval filter, a visual interface system for presenting multimodal contents, and wireless sensors. The goal of this work was to personalize learning materials for children with special needs. Figure 4(a) and 4(b) present, respectively, the overall architecture and the multimodal interface of the system described in [5].

![Figure 4. Architecture and multimodal interface of system presented in [5]](image)

Figure 4. Architecture and multimodal interface of system presented in [5]

The control of IoT sensors were built on a Raspberry Pi board (Figure 5) using a Linux-based operating system.

![Figure 5. Raspberry Pi board for controlling sensors (source [5])]
The purpose of work presented in [18] was to create a smart environment combining the IoT, P2P (Peer-to-Peer) communication, web-based and sensor technologies to monitor and control health status, and to assist children with ASD (Figure 6).

![Figure 6. Architectural model for the system presented in [18].](image)

The P2P system is based on a JXTA-Overlay Platform, responsible for communications between users. The SmartBox (Figure 7) is a hardware built to manage children’s monitoring sensors, among which we can highlight the sensor for hand and body moves, for detecting chair vibrations, for controlling ambient lighting, for controlling ambient odors and for reproducing audio. Moreover, by means of the Heuristic Diagnostic Teaching (HDT), this smart environment was used to identify mathematical learning skills and creative characteristics for students that are diagnosed with autism.

![Figure 7. SmartBox (source [18]).](image)
The last article found in the search [16], proposes a gesture and object recognition system for mobile platforms with the Android operating system, named MOBIS, which can assist autistic children to recognize objects when doing learning activities (Figure 8). This system captures images by means of a cell phone or tablet camera and processes them by using visual support techniques. The system uses a vision-based object recognition algorithm to associate visual and oral markers to the object that is being captured and recognized. It associates the images of the objects to other textual and sound elements, in addition to other equivalent images found on the Internet.

![Figure 8. MOBIS system for object visual recognition (source [16]).](image)

Table 2 show a summary with the main technologies used in the selected articles. These technologies were classified in software, hardware and communication technologies. Hardware technologies are mostly related to the sensors, actuators and devices identified in the selected articles.

<table>
<thead>
<tr>
<th>Software</th>
<th>Hardware</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>- machine learning algorithm</td>
<td>- audio generation</td>
<td>- P2P Communication</td>
</tr>
<tr>
<td>- web information retrieval</td>
<td>- sensor for controlling odors</td>
<td>protocol</td>
</tr>
<tr>
<td>filter</td>
<td>- sensor for controlling lighting</td>
<td>- Internet</td>
</tr>
<tr>
<td>- visual interface system for</td>
<td>- movement sensors</td>
<td>- Wireless</td>
</tr>
<tr>
<td>presenting multimodal contents</td>
<td>- Raspberry Pi board</td>
<td></td>
</tr>
<tr>
<td>- monitoring sensors (temperature, vibration, motion)</td>
<td>- mobile devices (cell phones and tablets)</td>
<td></td>
</tr>
<tr>
<td>- image processing and visual computing</td>
<td>- use of tags and RFID</td>
<td></td>
</tr>
</tbody>
</table>
4. SCALA and the Internet of Things

This section addresses the proposal on how to integrate IoT technologies with the SCALA System. The initial module to support IoT in SCALA will be the Communication Board module. Table 3 shows a comparison between the functionalities of the Communication Board module of SCALA and the most used technologies resulting from the exploratory research (Table 2).

<table>
<thead>
<tr>
<th>SCALA functionalities</th>
<th>IoT technologies applied with Autism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Board layout options</td>
<td>1. machine learning algorithm</td>
</tr>
<tr>
<td>2. Create board</td>
<td>2. web information retrieval filter</td>
</tr>
<tr>
<td>3. View board</td>
<td>3. visual interface system</td>
</tr>
<tr>
<td>4. Undo command</td>
<td>4. monitoring sensors</td>
</tr>
<tr>
<td>5. Edit image</td>
<td>5. audio</td>
</tr>
<tr>
<td>6. Save board</td>
<td>6. odor sensors</td>
</tr>
<tr>
<td>7. Open board</td>
<td>7. lighting sensors</td>
</tr>
<tr>
<td>8. Add pages to board and clear board</td>
<td>8. Raspberry Pi board</td>
</tr>
<tr>
<td>9. Export board</td>
<td>9. communication protocol</td>
</tr>
<tr>
<td>10. Import images</td>
<td>10. P2P</td>
</tr>
<tr>
<td>11. Print board</td>
<td>11. Internet</td>
</tr>
<tr>
<td>13. Free Communication module</td>
<td>13. mobile device</td>
</tr>
<tr>
<td></td>
<td>14. visual computing</td>
</tr>
<tr>
<td></td>
<td>15. tags and RFID</td>
</tr>
</tbody>
</table>

There are 13 functionalities in the SCALA System and 15 main technologies identified on articles about IoT and Autism. Some of these technologies already are being used in SCALA. The Communication Board module has already incorporated visual interface, audio generation, support for Internet and mobile devices technologies.

But, in order to allow a full support of IoT on SCALA the following road-map is highlighted: (1) Incorporate the identification of tags/RFID in physical objects; (2) Implement visual recognition algorithms for physical objects; (3) Adapt the Free Communication module in order to use pictographs represented by physical objects; (4) Elaborate a mapping strategy, between physical objects with virtual objects (pictographs). Web information retrieval systems and semantic technologies may be used to identify possible models of equivalent pictographs; (5) Encapsulate sensors in physical objects to identify approximation, location and touch, to execute audio, to show lights, for vibration, among other features; (6) Develop a routing protocol for a local network of physical objects, supporting on information security. In this case, a protocol like SPTP [1], which uses a secure P2P communication technology, can be used.

The steps proposed above define our next goals in respect to the evolution of SCALA system on the IoT. We expect that the fulfilling of these goals will make the experience to use the SCALA alternative communication system more productive and fun, bringing positive pedagogical results.
5. Final Remarks

The exploratory research methodology used to find works that apply Internet of Things technologies to Autism was effective because descriptors limited the scope of the research. The individual analysis of the each one of the 13 articles initially found, allowed us to identify three recent works with good ideas on how to proceed with the integration of IoT with SCALA. The detailed analysis of these works resulted on 15 examples of technologies that could then be compared to the functional characteristics already present on the SCALA system. The entire analysis and comparing process allowed us to propose a set of six new technological features to be incrementally developed and added to SCALA, resulting, in the end, in a system able to handle and transparently relate virtual and physical objects for a more productive alternative communication experience.

6. Acknowledgements

The authors are grateful to Foundation for Research Support of the State of Amazonas - FAPEAM (by means of RH-DOUTORADO Scholarships), Federal Education Institute of Amazonas (IFAM), Federal University of Rio Grande do Sul (UFRGS) and University of Vale do Rio dos Sinos the financial support given to this research.

References


Model of assessment using tests formed of questions with different theoretical and practical difficulty degrees based on a genetic algorithm

Doru Popescu Anastasiu¹, Nicolae Bold²

¹University of Pitesti, Faculty of Mathematics and Computer Science, Romania, dopopan@gmail.com
²University of Agronomic Sciences and Veterinary Medicine Bucharest, Faculty of Management, Economic Engineering in Agriculture and Rural Development, Slatina Brunch, Romania, bold_nicolae@yahoo.com

Abstract: Assessment is an important process in education, because it provides feedback from the assessed persons. Through this process, the assessor discovers the degree of knowledge of the taught notions. In the context of the development of technology, models of assessment diversified or were updated using technology means. This paper shows a type of model of generating tests using questions with different degrees of difficulty, given a desired interval of difficulty. This model is based on an algorithm which uses genetic notions. This model has several domains of applicability, one of the main being assessment within education. Also, it consists in a structure of a future web application which will use technology-based means for the assessment of students and which will be described in detail in the paper.

Keywords: test, genetic algorithm, difficulty, assessment.

1 Introduction

Generating assessment tests with the characteristics of equitability and efficiency for both the assessor and the assessed persons can be a challenging task. Moreover, the high number of questions which consist in the question database for a test is a serious time-related obstacle for the selection of the most adequate questions for a specific situation.

These reasons have led to the creation of a model called The Dynamic Model of Assessment and Interpretation of Results (DMAIR) which would encapsulate all this requests in a single web or offline application.

Diverse models for generating specific tests were presented in literature. The methods presented in these papers are different, using evolution-based algorithms, e.g., genetic and memetic algorithms (such as in papers [2] and [3]), combined with divide-and-conquer elements, such as in [1] or just plain generation, such as in [5]. Other cases are studied in the literature, such as generation of tests taking into account the notion of concept-effect model [4].

This paper introduces concepts such as theoretic difficulty degree (TDD) or practical difficulty degree (PDD). Based on these coefficients, a degree of difficulty (DD) of a question is calculated using a formula which will be presented in section 3.
2 Description of the Dynamic Model for Assessment and Interpretation of Results (DMAIR)

This model of assessment is part of a larger system which reunites database-related elements, graphic user interfaces, generation models and mechanisms of data interpretation and analysis. This system is intended to be used with the scope of the assessment of the students. The general scheme of this model is presented in Figure 1.

Fig. 1: The DMAIR assessment model proposed by the authors

The most changeable elements within this model are the conditions. In previous papers, we made implementation of algorithms that treat various conditions. For example, paper [6] presents an implementation based on trees. Paper [7] shows an implementation and results for genetic algorithms where tests/questions are isolated.

3 Description of the method based on a genetic algorithm

Given a question \( q \), whose number of given answers is \( k \) and number of correct answers is \( k' \), the degree of difficulty is \( k'/k \). The degree of difficulty can also be approximated by the teacher.

The theoretical difficulty degree (TDD) of \( q \) is considered as being the degree in which \( q \) is considered difficult from a theoretical point of view (memorization etc.). The practical difficulty degree (PDD) of \( q \) is considered as being the degree in which \( q \) is considered difficult from a practical point of view (usage of thinking, analysis, data interpretations etc.) For a specific question \( q \), the degree of difficulty (DD) based on TDD and PDD is:

\[
DD = \frac{(TDD \times \alpha) + (PDD \times \beta)}{\alpha + \beta}
\]  
(1)
where \(0 \leq \alpha \leq 1, 0 \leq \beta \leq 1\) are proportions given by the user and symbolize: \(\alpha\) is the theoretical proportion given by the user, which determines the proportion in which the theoretical questions will be found in the test; \(\beta\) is the practical proportion, also given by the user, determines the proportion of practical questions in the test. The sum of these two coefficients can be 1 or less than 1, depending on the needs of the user.

Intuitively, the larger the theoretical coefficient (\(\alpha\)) will be, the larger the proportion of the theory-based will be. Analogous, a higher value of the practical coefficient (\(\beta\)) will determine the generation of more practical problems. The degree of difficulty of a test is the sum of all the DDs of the questions that form a test.

\[
DD_{test} = \sum DD\ of\ all\ questions\ within\ a\ test\ (2)
\]

Genes will be considered questions within a test and chromosomes will be considered the tests themselves. As genetic operations, we will use mutation and crossover with one point. The fitness value is represented by DD_{test} for each chromosome.

4 Implementation and results

Taking into account the viability, we took a generated test using the algorithm and we gave it to a number of 20 students to solve it. Because of the low development of the GUI, we used Google Forms for the answer to the questions. The number of questions in the database was set to 400 (N=400) and the test had 20 questions (M=20). The assessor wanted to test some theoretical issues, so \(\alpha = 0.9\) and \(\beta = 0.1\). A random sequence was chosen, that is the test contained the questions with order numbers (97 145 126 32 325 319 44 337 300 61 95 70 361 183 208 264 42 240 98 283). The degree of difficulty of this test was resulted at 13.99.

<table>
<thead>
<tr>
<th>Q</th>
<th>1-10</th>
<th>97</th>
<th>145</th>
<th>126</th>
<th>32</th>
<th>325</th>
<th>319</th>
<th>44</th>
<th>337</th>
<th>300</th>
<th>61</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>0.26</td>
<td>0.34</td>
<td>0.66</td>
<td>0.66</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>Q</td>
<td>11-20</td>
<td>95</td>
<td>70</td>
<td>361</td>
<td>183</td>
<td>208</td>
<td>264</td>
<td>42</td>
<td>240</td>
<td>98</td>
<td>283</td>
</tr>
<tr>
<td>DD</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 1: Degrees of difficulty for the sequence taken as example

The test was indeed 90% with theoretical issues. Only the first 2 questions requested solving some practical problems. The comparison of DD with the Difficulty Index (DI) is made in Figure 2.
Fig. 2: DD compared with DI

The behavior of the two axes is somehow normal. DD must have an increasing slope, because the greater the DD, the harder the question, while the less the value of the DI, the greater the difficulty of the question. Thus, the DI axis must have a decreasing slope. Here, we must differentiate between practical and theoretical questions. For example, if $\alpha$ is greater, than a practical question can be seen as an “easy” question, even if its DD can be bigger (the question is “hard”) when $\beta$ is bigger. The complexity of the algorithm is $O(NG\timesNrPop\times M)$. In terms of runtime, for $N=400$, $M=100$ and $NG=1000$, the algorithm had a runtime of 14.05 seconds.

5 Conclusions

The DMAIR is a model that use technology means to ease the job of the assessor. Sometimes, selecting questions that fit certain requirements can be a time-and-energy consuming task. Thus, this model increases the efficiency of this process of education, by emphasizing the difference between theoretical and practical aspects of assessment. In these terms, efficiency is seen as obtaining tests with the highest fitness values which respect requirements related to increasing theoretical and practical degrees of difficulty. As future work, we would like to implement the model using past research, giving some improvements to other mechanisms within the model.

6 References


The Impact of Social Similarities and Event Detection on Ranking Resources in Collaborative E-learning

Samia Beldjoudi¹², Hassina Seridi², Abdallah Benzine²
¹ Preparatory school of sciences and technology Annaba, Algeria
² Laboratory of Electronic Document Management LabGED, Badji Mokhtar University
Annaba, Algeria
{beldjoudi, seridi}@labged.net
abdoulahbenzine@gmail.com

Abstract. Recently, folksonomies have become a very popular method to describe social web contents due to their simplicity of use. The objective of this paper is to personalize tag-based search within E-learning folksonomy and extract implicitly the semantics of learners’ tags in order to overcome the ambiguity problem. On another side, we want to study whether exploiting other features together with event detection can help improving the retrieval effectiveness when searching collaborative applications. In this context, we propose capturing the similarity between an information need and personalized resources to combine them with event dimension for relevance ranking. The aim is to acquire appropriate resources inquired by learners.

Keywords. Social Tagging, Folksonomies, E-learning, Event detection, Ranking.

1 Introduction

In social web, users create, annotate, share and make public what they find interesting on the web. Folksonomies are one of these new social practices that can be considered as classification systems derived from collaborative practice allowing tags’ creation and management in order to annotate and categorize content. Recently, E-learning platforms focused on personalization to achieve some pedagogic scenarios otherwise inaccessible in traditional learning. The aim of this work is to focus on resources personalization to learners when they search relevant resources by use of tags. Unfortunately, freely chosen tags in folksonomies are likely to contain spelling errors and therefore make the resources retrieval more doubtful than the metadata recovering from a lexicon examined by domain professionals. Also, a large number of resources can match users’ queries in folksonomies. Therefore, ranking these web resources is required to release user who cannot browse all them. In this work, we focus on resource retrieval in folksonomies where we intend to overcome tags’ ambiguity. Further we propose a ranking function to rank resources where we combined two features: social aspect and event detection. The rest of this paper is organized as follows: Section 2 describe the proposed approach. Section 3 presents some experimental results to measure the approach performance. Conclusions and future works are discussed in Section 4.

© Springer-Verlag Berlin Heidelberg 2016
2 Improving Tag-based Search in Collaborative E-learning Environment

In E-learning applications, a folksonomy is a tuple $F = \langle L, T, R, A \rangle$ where $L$, $T$ and $R$ represent respectively the set of learners, tags and resources, and $A$ represents the relationship between the three elements i.e. $A \subseteq L \times T \times R$. We consider a folksonomy as being a tripartite model where resources are associated with a learner ($l$) to a list of tags. Therefore we have extracted three social networks represented by three matrices $LT$ (learner-tag), $RT$ (resource-tag), $RL$ (resource-learner).

In these systems, users face some other problems like tags ambiguity. In our previous works (Beldjoudi et al., 2011, 2012), we proposed a method to treat the tag ambiguity problem based on computing social similarities between users. The approach takes benefits from the strength of community effect that characterizes Web 2.0 technologies. The main idea was to analyze each user profile, and then compares his preferences with other users in order to extract those who are similar to him. In this work, we want to see the impact of the proposed approach on searching resources rather than a recommender process, also we propose a ranking function to rank the retrieved resources based on social aspect and event detection features. The approach is tested within a collaborative E-learning application. Let us present the following illustrative example to give more details about the proposed approach.

![Illustrative Example](image)

**Fig. 1.** An illustrative example.

As presented in Figure 1, the learner $L_1$ want to obtain relevant resources related to the topic ‘Sun’. It is clear that before having a look at the learner's profile, we cannot know his preferences because the used tag is ambiguous. For a human reading $L_1$’s tags, we can conclude that $L_1$’s preferences are related to computer and not to biology field. To formalize this approach, we proceed as follow:

For each submitted query, we measure the similarities between this user and the users of his social network who use resources annotated with the tags occurring in this query. To measure similarity between two learners ($L_1$ and $L_2$), both are represented by a binary vector representing all their tags (extracted from matrix $LT$) and we compute the angle cosines between the two vectors.

To avoid problems that generally result from a lack of data required by the system in order to make a good recommendation, when the user of the recommender system is not yet similar to other users, we propose to measure similarity between resource.

In another side, when querying folksonomies, a large number of resources can match users’ queries. Therefore, ranking these web resources is a key problem. A ranking function should incorporate many features to be effective (features of the document, the query, the user ...) (Bouadjene et al., 2013). In this approach, a personalized social ranking function is proposed to provide personalized retrieved resources.
Let's consider an E-learning folksonomy \( F \) (L, T, and R) from which a learner \( l \in L \) submits a query \( t \) to the search engine. We would like to rank the set of resources that match \( t \), such that relevant resources for \( l \) are highlighted and pushed to the top to maximize his satisfaction and personalize the search results.

The objective of this approach is to improve resources retrieval and exploit the dependence between event detection and the high number of similar queries transmitted in the same period. The presence of an event \( e \) involves increasing the number of searches performed on this event. To achieve a good ranking, we must sort the retrieved resources according to some criterion so that the most relevant results appear early in the retrieved list displayed to the learner. Two cases can be observed:

1) There is no event detected during the search phase: in this case the retrieved resources will be ranked only according to social similarities values.

2) Otherwise, the proposed ranking function incorporated two features to be effective: the social aspect and the visit number of each resource during the same time period of event detection.

Thus the proposed function \( \text{Rank} \ (r, t, l) \) is computed by merging the average similarities values \( \frac{\sum_{j=1}^{n} \text{sim} (t, j, l)}{j} \) and \( \text{Nbr\_visit} \ (r) \). This merge is computed as follows:

\[
\text{Rank} \ (r, t, l) = \alpha \times \frac{\sum_{j=1}^{n} \text{sim} (t, j, l)}{j} + (1 - \alpha) \times \text{Nbr\_visit} \ (r) \quad (1)
\]

Where, the parameter \( \alpha \) denotes the weight that satisfies \( 0 \leq \alpha \leq 1 \) and represents the importance that we want to give to the features’ types (i.e, social similarities or most popular resource in \( n \) time units). Note that, the first side of the formula (1) represents the average value of similarities between \( l \) and the \( j \) other learners who tagged a given resource \( r \) with the tag \( t \). Introducing the last feature in the ranking function is crucial, because the presence of an event can have a real influence on the resources popularity and thus have an impact on the query meaning. Examples of events that can affect the learners search are: Exams, TPs, Reports, BAC, etc. where learners can submit similar queries in these periods.

3 Experiments and Evaluation

An experiment over del.icio.us dataset is conducted; we were interested with data sample constructed from users who tagged resources about education. Our data base comprises 20432 tag assignments involving 7898 users, 15439 tags some of them are ambiguous, 10527 resources each having possibly several tags and several users. Note that, the used dataset include also the date of each tagging operation, this can help us in event detection. To evaluate the proposed approach, we used the metrics: recall, precision and F1 metric. The three metrics are calculated to each user, and then the average of each metric in the system is calculated. The results are shown in Table 1:

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous Tags</td>
<td>85%</td>
<td>80%</td>
<td>83%</td>
</tr>
<tr>
<td>Not ambiguous</td>
<td>93%</td>
<td>85%</td>
<td>89%</td>
</tr>
</tbody>
</table>

From the analysis of the above table, we can conclude that the Precision, Recall and the metric F1 in the approach are very promising both in the case of ambiguous tag-
based queries and also not ambiguous tag-based queries. This result indicates that social similarities and event detection performed by our approach are really able to help learners when they query a folksonomy.

To evaluate the impact of these events on ranking the retrieved resources, we proceeded as follow: When we detected the presence of an event, two experimental scenarios to choose the value of $\alpha$ parameter were proposed:

1. **The $\alpha$ value is between 0 and 0.4 ($0 \leq \alpha \leq 0.4$)**

   In this case, the ranking function is focused much more on the second feature of the formula (1), which is the popularity of resources when event detection. Experiments demonstrated that in the case of ambiguous tags, when two different events are detected, precision value is bended down because we have neglected the social similarities between users that can personalized the retrieved results. Thus focusing only on event detection can decrease the pertinence of ranking function.

2. **The $\alpha$ value is between 0.5 and 1 ($0.5 \leq \alpha < 1$)**

   In this second case, the ranking function focused on the social aspect more than the new event influence on resource popularity. This can degrade the result precision because even with the fact that users are similar, there are some resources becoming not relevant to an active learner because they are very former and thus their popularity decreased comparatively with new resources that are presented recently with event detection. Experiments demonstrate that the most suitable value of $\alpha$ that produces the highest value of precision, recall and F1 metric is 0.6. This implies that it is the most adequate when the user wants to obtain a trade-off between social similarities and new events detection.

4 Conclusion and Future Works

Investigations in the field of collaborative E-learning allowed us to demonstrate the impact of social similarities and event detection on ranking resources in collaborative E-learning. We tested the approach on a baseline dataset and we obtained promising results. In order to continue and improve the proposed work, we aim to carry some additional tests on other datasets, perform comparison with other ranking functions and also use the map-reduce framework in order to let the approach scale with very large databases.

5 References

Towards a Conceptual Framework for Bridging Player Roles and Learner Roles in Gamified Collaborative Learning Contexts

Simone S. Borges¹, Riichiro Mizoguchi², Vinicius H. S. Durelli¹, Ig. I. Bittencourt³, Seiji Isotani³

¹University of São Paulo, ICMC, São Carlos, SP, Brazil
{sborges, durelli, sisotani}@icmc.usp.br
²Japan Advanced Institute of Science and Technology, Ishikawa, Japan
mizo@jaist.ac.jp
³Federal University of Alagoas, Maceió, Brazil
ig.ibert@ic.ufal.br

Abstract. Gamification has been widely used in education with mixed results. Some empirical findings indicate that gamification can increase motivation and engagement in students; other findings highlight that gamification can be a distraction to students; and therefore it may end up hindering learning. To deal with this problem, it is necessary to design gamification to fit properly in individual and collaborative learning contexts. Unfortunately, there is a lack of studies on frameworks mapping game elements and learning theories to support the adequate design and application of gamification in education. Thus, the main goal of this paper is to present a conceptual framework based on collaborative learning theories and gamification literature that creates a link between player roles and learner roles. To demonstrate the viability of our framework to design theory-based gamified collaborative learning scenarios, we describe a case study using the theory Peer Tutoring.

1 Introduction

The term gamification originated in the digital media industry, however, such term only gained widespread acceptance after the end of 2010 [1]. Gamification refers to the use of game design elements like mechanics, aesthetics, and game thinking to non-game contexts that aim to engage people, motivate action, improve learning, and foster problem solving [2]. In gamification approaches, these game design techniques are not the center of the system. Rather, the aim of these gamification approaches is to improve the user experience. Given that applying game elements to education has the potential to maximize the engagement of students and motivate them [2], there has been a growing interest in applying gamification to education [3].

However, simply inserting game elements in a system and hoping for the best will not improve the system utilization [4]. The same applies to gamification initiatives that rely only on distributing points and badges (e.g., pointification, exploitationware, shallow gamification) [5, 6]. Therefore, building sound educational systems that rely...
on gamification techniques require careful analysis of the most suitable game design elements that will help to achieve the desired learning outcomes [2]. In this sense, the representation of learner roles and player roles, which rely on sound pedagogical basis, is highly desirable. Despite the well-known benefits of Computer-Supported Collaborative Learning (CSCL), collaboration in education only has a positive influence on the learning process if its design and utilization occurs in a proper way [7]. The carefully design of CSCL environments and collaborative learning (CL) activities is a vital task; otherwise, there is no guarantee that the learning outcomes will meet the desired expectations. In this context, the role played by group formation is fundamental: it influences how students perceive the environment, interact with peers, use available didactic materials, and take part in learning activities and processes [8].

Several researchers have been investigating how to design effective group formation methods to improve collaborative learning sessions. However, as pointed out in [6], even with the success of design scripts to support CL tasks, there are situations in which such scripts may interfere, at least to some degree, with students’ motivation. Hence, among other solutions, some researchers and practitioners have been investigating gamification as an alternative to engage and motivate students in CL scenarios [2, 9]. Since gamification is highly context-dependent, ill-designed gamification solutions can lead to harmful effects instead of the expected benefits [10]. Therefore, the construction of sound gamified CL sessions requires not only, careful analysis of appropriate gamified activities (e.g., environment’s design) and meaningful rewards (e.g., suitable game elements) but also, the ability to designate appropriate players’ roles for each learner. As a first step towards these objectives, we set out to create models and vocabulary to represent learner-player roles interactions. Nevertheless, it is worth emphasizing that the vocabulary and player roles presented here are by no means exhaustive.

In this paper, we present a subset of models we devised as well as the vocabulary gathered and organized to represent learner-player roles interactions in gamified CL contexts. In Section 2, we present related work. In Section 3, we discuss several topics that support our research. In Section 4, we introduce the methodology we used, the models, and a non-exhaustive example of vocabulary. In Section 5, we provide an example of how we intend to use these models and vocabulary, by describing a personalized learner-player role interaction in a Peer Tutoring scenario. Finally, in Section 6, we present concluding remarks and future research directions.

## 2 Related Work

Currently, to the best of our knowledge, no research effort has proposed a link between player roles to learner roles in CL. In [11], they present an e-learning environment wherein generic mechanics and player types found in the gamification literature were incorporated. They also investigated the effects of such mechanics into different learning activities aimed to observe the learning effectiveness of the selected mechanics and the relationship between the mechanics and player types. In [12], a gamified collaborative learning environment was built to foster group collaboration: the authors
applied gamification techniques to motivate a group of learners while they performed a task in a collaborative fashion. During this investigation, the authors employed visible progress of task completion as feedback. Also, in [6], an ontology to gamify CL scenarios is described, although the focus of the study is the formalization of basic concepts concerned to gamification as a persuasive tool in CL, the relationship between player roles and suitable learner roles are not taken into account.

Furthermore, despite many attempts to unify players’ behaviors and, as a result, establish an ultimate “players’ type taxonomy”, few studies have investigated players’ types from the standpoint of educational theories. There is a lack of studies investigating what kind of player role a learner should play to trigger the desired behaviors in gamified CL settings.

3  Background

Our approach for creating player roles is based on the motivations to play proposed by [13]. In [13], differently of most available player models, instead of using psychological archetypes in an effort to fit a player in one kind of dominant personality type, the proposed approach tries to identify not only the reasons that can motivate an individual to play a video game, but also their relationship and overlaps. In this way, the model proposed by [13] does not attempt to identify in which archetype one individual fits, but rather to understand what can motivate such an individual to play. In addition, by identifying the reasons that may arouse the motivation to play, by analyzing the results of the scoring system developed, one can also identify what can be deemed less attractive for such an individual.

To help us better understand player’s needs that drive motivations to play, we rely on Self-determination theory (SDT) concepts [14]. SDT seeks to explain how intrinsic and extrinsic motivators influence human behavior and the development of individuals. The three basic psychological needs, considered fundamental to influence motivation are autonomy, relatedness, and competence. According to [14], by promoting the internalization of these feelings, individuals have the potential to carry out their activities with improved performance, persistence, and creativity, for instance.

4  Method

As a first step, we designed a protocol to collect and analyze the literature on group formation in CSCL. More specifically, we carried out a systematic mapping study by following the guidelines proposed by [15]. Initially, we collected 3571 papers about CSCL that had the potential to provide valuable information about research on group formation. After a careful analysis of each paper, only 106 met the necessary requirements/criteria defined in our protocol. As a result, each of the 106 papers was categorized according to their contributions. As a second step, we performed another mapping study to investigate the state of the art of how gamification has been applied in educational settings. In addition, we studied the previous efforts developed by [8, 16–18]. These studies are based on sound instructional and learning theories and pre-
sent significant contributions on how to modeling and conducting CL activities in intelligent learning environments. The information extracted from both mapping studies gave us a useful overview of the state of the art of both domains, such as new methods, algorithms or criteria for group formation; the rationale (or the absence of one) behind the group formation strategy; and also possible advantages, drawbacks and pitfalls when using gamification in learning environments.

Next, we devised our definition of player roles crossing information extracted from (i) motivations to play, (ii) SDT concepts, and (iii) player’s types found in the gamification literature. Considering that our main goal was not to come up with a new player typology, we decided to keep things simple, by initially limiting the number of player’s types to five. Although this is not an exhaustive set of players, we believe that this granularity can cover a reasonable scope of the most common player types found in the gamification literature, therefore fulfilling the purpose of the definition of our player roles. After this step, we compared the chosen player’s types with motivations to play, and psychological needs (extracted from SDT).

As shown in Table 1, we linked the components Achievement, Social, and Immersion, with the following psychological needs: Competence, Relatedness, and Autonomy, respectively. The component Achievement has the subcomponents: Advancement, Mechanics, and Competition. In order to keep things simple, as shown in Table 1, some subcomponents and psychological needs were grouped together according to their synergies. We related the subcomponents Advancement and Mechanics to the psychological need Competence. We decided to join them because we consider that they are connected to the player type; the Achiever. To support this decision we checked many player typologies [19–22] looking for information that could help us devise appropriate player roles for each subcomponent. We found, for example, that in gamification literature, Achievers enjoy not only beating a game but also being the most successful player, accumulating rewards in the process. To do this, achievers strive to understand the game mechanics and/or they scrutinize the game to reach their goals. In other words, the achiever behavior can be seen as goal-oriented, and people experiencing this behavior will expend a lot of time to reach their goals. We used the same approach to connect the remaining components, subcomponents, psychological needs and player types.

In Table 1, both Achievers and Killers belongs to the component Achievement, however, while Achievers are more interested in increasing their in-game reputation by completing different tasks, Killers are more interested in tasks that involve besting other people in some sense. In other words, Killers are people-oriented, and they score high in the subcomponent Competition. Killers, in game design literature, are the type of players highly motivated by Competition. They find zero-sum game mechanics appealing [21, 23], so they enjoy rushing and competing against other people.

Next, we grouped all social subcomponents in a single player type, Socializer. Socializers are motivated mainly by interacting with people (e.g., being in groups and teams, forming partnerships, and playing collaborative games). Like Killers, Socializers are also people-oriented, however, instead of emphasizing defeating other players, they value socializing, sharing experiences, building relationships, and performing shared tasks. Finally, Explorers were related to the subcomponent Discovery. Explor-
ers are system-oriented players, and they are motivated by exploration: these players are drawn to the possibility of investigating the system’s ins-and-outs (e.g., hidden or remote places, finding loopholes, knowing the rules that govern a space). Creators were linked to Customization: these players are also system-oriented, but they are more interested in customizing the system or modifying the virtual world (e.g., backgrounds, fonts, buildings, characters, weapons, and vehicles).

Table 1. Motivations to play X Self-determination Theory X Player types

<table>
<thead>
<tr>
<th>Component</th>
<th>Subcomponent</th>
<th>Psychological needs</th>
<th>Player Types [21]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Competence Relation</td>
<td>Achiever</td>
</tr>
<tr>
<td>Achievement</td>
<td>Advancement</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td>Socializing</td>
<td>●</td>
<td>Socializer</td>
</tr>
<tr>
<td>Social</td>
<td>Relationship</td>
<td>●</td>
<td>Explorer</td>
</tr>
<tr>
<td>Social</td>
<td>Teamwork</td>
<td>●</td>
<td>Creator</td>
</tr>
<tr>
<td>Immersion</td>
<td>Discovery</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Immersion</td>
<td>Customization</td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>

Once we gathered all information in Table 1, we devised the five player roles shown in Figure 1 and described in Table 2. Although we borrow all names from the gamification literature, it is important to note that, instead of player types, we are looking for the creation of a set of player roles. According to [24], although there is no consensus about the definition of roles, it is possible to assume that a role is an entity that is played by another entity in a context. In this sense, “context” can be understood as something as a whole, including a relation in which the former “entity” is defined [24].

Fig 1. Player roles [6] [8]

Considering these player roles are meant to be used in educational environments, we carefully chose these labels, among several labels found in the literature, to avoid possible negative connotations (e.g. killer, exploiter). As pointed by [23] and [25], to avoid the externalization of bad and/or undesirably behaviors, it is important not to provide the system with the kind of game design element that can lead to harmful behavior. As an example, Blizzard’s electronic card game Hearthstone® does not allow players to chat with each other. There are pre-selected emojis that can vary slightly depending on the character. However, there is no way to send your custom
text message to your opponent. The company recognizes the huge tradeoffs involved in such decision, however, since the game can be very competitive, they stated that this was the right decision to keep Hearthstone® fun, safe, and appealing to most players [26]. By substituting the chat system to pre-selected emotes, Blizzard tailored their game in order to restrain inappropriate vocabulary and harassment.

Moreover, although one can argue that since the research focus is on collaborative learning activities, why players such as Achiever and Conqueror should be considered, given that their specific nature is not collaborative at all. The answer is that all depends on the context. The structure of the interaction between players is a choice to be decided early in any game project (or gamified project). Although the most well-known interaction pattern is the player-vs-player, there are others patterns to be considered [27]. Community collaboration pattern, for example, can be used to pose challenges to the users to work against the system in a collaborative fashion (e.g., tasks related to time limit or that have to be performed against an Artificial Intelligence).

Table 2. Player roles description

<table>
<thead>
<tr>
<th>Player role</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achiever (goal-oriented)</td>
<td>They enjoy not only completing a game, but also being the best winner, accumulating all rewards the game can offer. They are motivated by receiving glory (points, titles, medals, trophies, achievements); gathering (virtual currency and goods); and collecting rare or all in-game items (equipment, weapons, armor, vehicles, mounts, pets).</td>
<td>[13, 19, 21, 23]</td>
</tr>
<tr>
<td>Conqueror (people-oriented)</td>
<td>They enjoy rushing and competing against other people. Usually, they enjoy testing their skills and seeing how they stack up against other people. They find external ranking systems and zero-sum game mechanics appealing.</td>
<td>[13, 19, 22, 23]</td>
</tr>
<tr>
<td>Humanist (people-oriented)</td>
<td>They enjoy socializing with people (e.g., being part of groups and teams, forming partnerships, and playing collaborative games). They value socializing, sharing learning, and relationship building via shared tasks.</td>
<td>[13, 19, 22, 23]</td>
</tr>
<tr>
<td>Explorer (system-oriented)</td>
<td>They enjoy exploring the system by discovering the ins-and-outs (e.g., hidden or remote places, finding loopholes, knowing the rules that govern a space).</td>
<td>[13, 19, 22, 23]</td>
</tr>
<tr>
<td>Creator (system-oriented)</td>
<td>They enjoy customizing the system (e.g., backgrounds, fonts, buildings, characters, armor, weapons, and vehicles).</td>
<td>[13, 19, 21, 23]</td>
</tr>
</tbody>
</table>

5 Vocabulary to Represent Interactions in Gamified Collaborative Learning

In Table 3, we present a non-exhaustive compilation of behaviors and player’s types. It is important to remember that, as stated before, there is no guarantee that all behaviors can be useful or appropriate for an educational environment, therefore behaviors should be carefully chosen to avoid triggering undesirable interactions. The information presented in Table 3 should be read as, for example, “The Achiever is
acting on the System by Comparing his Progress (e.g., on the leaderboard)”. We used such information as a starting point to devise appropriate behaviors that one can use in a gamified CL environment to improve learning. By appropriate we mean behavior that is able to motivate learners to play and satisfies the respective psychological needs established by SDT. Due to space constraints, in Table 4 we will present only a subset of selected behaviors, one behavior for each player role. As any model, the scheme presented in Table 3 is a simplification of the reality and, as such, it has some drawbacks.

Table 3. Examples of player types and common behaviors found in the game-related literature [13, 19, 21, 23, 28]

<table>
<thead>
<tr>
<th>Player types</th>
<th>Performance</th>
<th>Who/what</th>
<th>Doing (behavior)</th>
<th>who/what</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achiever</td>
<td>acting</td>
<td>System</td>
<td>Tracking, Collecting, Gathering, Pursuing, Comparing, Bragging, Showing-off, Winning</td>
<td>Progress, point, medal, achievement, honor, currency, items, virtual goods, quests, status, leaderboard, prizes</td>
</tr>
<tr>
<td>Explorer</td>
<td>interacting</td>
<td>System</td>
<td>Discovering, Exploring, Viewing, Investigating, Traveling, Searching, Looking for</td>
<td>Virtual world, Maps, remote places, secret places, Easter-eggs, hidden quests</td>
</tr>
<tr>
<td>The Killer</td>
<td>in acting</td>
<td>Others</td>
<td>Harassing, Hacking, Killing, Disturbing, Troublemaking, Defying, Cheating, Taunting, Teasing, Fighting</td>
<td>Players, characters, GMs (game masters) NPCs (non-player characters), guilds</td>
</tr>
<tr>
<td>Socializer</td>
<td>interacting</td>
<td>Others</td>
<td>Helping, Greeting, Giving, Supporting, Sharing, Collaborating, Commenting</td>
<td>Players, characters, GMs (game masters) NPCs (non-player characters), information, forums, guilds</td>
</tr>
<tr>
<td>Creator</td>
<td>interacting</td>
<td>System</td>
<td>Creating, Tweaking, Building, Customizing, Transforming, Adapting, Inventing, Crafting</td>
<td>Interface, maps, MODs (modification), avatar, weapons, armors, vehicles and mounts</td>
</tr>
</tbody>
</table>

We will now demonstrate the viability of our framework. In our example, we will use the Peer Tutoring theory. Peer tutoring considers that both tutor and tutee can gain benefits from their interactions. Essentially, the tutor is a more knowledgeable individual and the tutee is a less knowledgeable individual. However, tutor and tutee are learners who do not have the “complete” knowledge about the content. Therefore, by facing difficulties in teaching, the tutor needs to acquire more knowledge in order to teach and organize his thoughts in an understandable manner; and through the tutoring process, the tutee will acquire or construct his/her knowledge as well.
Table 4. Player roles and potential behaviors adapted from the literature

<table>
<thead>
<tr>
<th>Player role</th>
<th>Behavior</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achiever</td>
<td>Tracking</td>
<td>To keep track of other learners progress and his/her own progress.</td>
<td>[13, 19, 22, 28, 29]</td>
</tr>
<tr>
<td>Humanist</td>
<td>Empathizing</td>
<td>To have empathy for other learners.</td>
<td>[13, 19, 22, 23]</td>
</tr>
<tr>
<td>Conqueror</td>
<td>Troublemaking</td>
<td>To unsettle the other learners by proposing solutions that are sometimes trustworthy but other times erroneous.</td>
<td>[13, 19, 22, 23]</td>
</tr>
<tr>
<td>Explorer</td>
<td>Exploring</td>
<td>To unveil new things in the system as the learning process progresses (e.g., features, tasks, content).</td>
<td>[13, 21, 28, 29]</td>
</tr>
<tr>
<td>Creator</td>
<td>Tweaking</td>
<td>To tweak the system towards personalizing something (e.g., learning path, tasks).</td>
<td>[13, 21, 28, 29]</td>
</tr>
</tbody>
</table>

Table 5. Peer tutor role, Player roles, Prerequisites, and Expected effects

<table>
<thead>
<tr>
<th>Learner Role</th>
<th>Player Role</th>
<th>Conditions</th>
<th>Expected effects</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achiever</td>
<td>* Knowing the goal state (target) that the tutees should reach.</td>
<td>(PL) Acquisition of the content specific knowledge by keeping track of tutees progress (tuning); (ML) Satisfaction of the need for Competence keeping track of tutees progress.</td>
<td>[13, 19, 22, 28, 29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Knowing the tutees’ confidence level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Having the pedagogical expertise to maximize the impact of its interventions.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conqueror</td>
<td>* Knowing the goal state (target) that the tutees should reach.</td>
<td>(PL) Acquisition of the content specific knowledge by unsettling the tutees (tuning); (ML) Satisfaction of the need for Competence unsettling the tutees.</td>
<td>[13, 22, 28, 29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Knowing the tutees’ confidence level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Feeling connected with the other tutees.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Knowing the tutees difficulties</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanist</td>
<td>* Knowing the goal state (target) that the tutees should reach.</td>
<td>(PL) Acquisition of the content specific knowledge by helping to solve the problems faced by the tutees (tuning); (ML) Satisfaction of the need for Relatedness though helping others to solve the proposed problems.</td>
<td>[13, 23, 28, 29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Feeling connected with the other tutees.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Knowing the tutees difficulties</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explorer</td>
<td>* Knowing the goal state (target) that the tutees should reach.</td>
<td>(PL) Acquisition of the content specific knowledge by progressively discovering new content while helping the tutees (tuning); (ML) Satisfaction of the need for Autonomy by having new experiences along with the tutees.</td>
<td>[13, 21, 28, 29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Knowing how to discover new content that the tutees should learn.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Not familiar with all system content.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Learner Role</th>
<th>Player Role</th>
<th>Conditions</th>
<th>Expected effects</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>* Knowing the goal state (target) that the tutees should reach.</td>
<td>(PL) Acquisition of the content specific knowledge by customizing the tutees' tasks (tuning); (ML) Satisfaction of the need for Autonomy by having choices.</td>
<td>[13, 21, 28, 29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Knowing how to choose which tasks the tutees should do.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Knowing the tutees difficulties</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explorer</td>
<td>* Knowing how to discover new content.</td>
<td>(PL) Acquisition of the content specific knowledge by progressively discovering new content (accretion); (ML) Satisfaction of the need for Autonomy by having new experiences.</td>
<td>[13, 21, 28, 29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>* Not familiar with all system content</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanist</td>
<td>* Feeling connected with the tutor.</td>
<td>(PL) Acquisition of the content specific knowledge by being assisted by the tutor (accretion); (ML) Satisfaction of the need for Relatedness through being helped to solve his/her problems.</td>
<td>[23, 28, 29]</td>
<td></td>
</tr>
<tr>
<td>Conqueror</td>
<td>* Pushing the tutor by expressing his/her ideas in a convincing manner.</td>
<td>(PL) Acquisition of the content specific knowledge by unsettling (flustering) the tutor (accretion); (ML) Satisfaction of the need for Competence by unsettling the tutor.</td>
<td>[22, 28, 30]</td>
<td></td>
</tr>
</tbody>
</table>

The first columns contain the planned Learner role and the Player role, respectively. In the column Conditions, we describe the necessary conditions (*) and the desired conditions (-) that a learner should fulfill to play both roles. The Expected effects column describes both learnings, and psychological outcomes, named pedagogical lenses (PL), and motivational lenses (ML). Assume two students, one playing the Peer tutor role, while the other plays the Peer tutee role. Both learners take a test to identify their players’ profiles. In addition, it is assumed that the results indicate that the learner playing the tutor role scored high as achiever, while the other student, playing the tutee role, scored high as a creator. Using information from Tables 5 and 6, we can combine this pairs of learners as follows:

Learner role: Peer tutor – Player role: achiever -> Tutor-Achiever role
Learner role: Peer tutee – Player role: creator -> Tutee-Creator role
In Table 7 we present both roles extracted from Table 5 and 6, respectively. We indicate the necessary and desired conditions in order to reach the hypothetical learner goals planned for both learners. Based on this information, designers and instructors can choose, more easily, suitable game design elements that will support the necessary and desired conditions.

<table>
<thead>
<tr>
<th>Learner Player</th>
<th>Role</th>
<th>Conditions</th>
<th>Expected effects</th>
<th>Game elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer Tutor</td>
<td>Achiever</td>
<td># Knowing the goal state (target) that the tutees should reach.</td>
<td>(PL) Acquisition of the content specific knowledge by keeping track of tutees progress (tuning); (ML) Satisfaction of the need for Competence keeping track of tutees progress.</td>
<td>Progress bars, points, medals, achievements, honor system, currency, virtual goods, quests, leaderboard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Knowing the goal state (target) that he/she should reach.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer tutee</td>
<td>Creator</td>
<td># Knowing how to choose tasks to do.</td>
<td>(PL) Acquisition of the content specific knowledge by being able to choose the necessary tasks to reach the learning goal (accr...</td>
<td>Custom interface, progress map, knowledge map, avatars</td>
</tr>
<tr>
<td></td>
<td></td>
<td># Tweaking the system along with the tutor.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 Concluding Remarks and Future Work

We outlined ways to relate player roles found in game design and gamification literature and possible ways to connect these player roles to CL learner roles in a systematic way. We posit that is possible to improve learners’ experience, by providing personalized situations, therefore contributing to minimize potentially harmful aspects of gamification, such as being disruptive or unattractive to some users. By identifying learners’ motivation to play, we can choose appropriate players’ roles for each learner in different gamified learning scenarios. Moreover, instead of classifying learners in a static manner, we devised roles that can be played in order to satisfy not only constraints related to the learning objectives but also the kind of constraints regarding the gamified CL environment.

Although some studies have already applied gamification in different cases, most efforts fail to provide a rationale to back up these research efforts. Furthermore, some studies do not rely on sound learning theories, relying only on ad hoc choices of both gamification techniques and learning scenarios. We presented one example of how our framework could be instantiated in a peer-tutoring scenario. We chose Peer learning as an example because it is a well-known (and effective [31]) theory, therefore minimizing our effort to explain both the learner roles and the player roles, as well as
their interplay. Due to space constraints, we did not show the other schemes we have developed for Anchored Instruction, Cognitive Apprenticeship, Cognitive Flexibility, among others learning theories and instructional strategies. Moreover, Table 4 does not show a complete list of investigated behaviors, also due to space constraints. Currently, we are working on the implementation of a prototype that will provide the necessary computational support to evaluate the effectiveness of our proposal.

Acknowledgments

The authors would like to thank CNPq and CAPES for supporting this research.

References

28. Inaba, A., Mizoguchi, R.: Learners’ roles and predictable educational benefits in collaborative learning an ontological approach to support design and analysis of


The Role of Agent-based Simulation in Education

Andrew Koster\textsuperscript{1}, Fernando Koch\textsuperscript{1}, Nicolas Assumpção\textsuperscript{2} and Tiago Primo\textsuperscript{1}

\textsuperscript{1} SAMSUNG Research Institute, Campinas, SP, Brazil \{andrew.k, fernando.koch, tiago.t\}@samsung.com
\textsuperscript{2} Instituto de Pesquisas Eldorado, Campinas, SP, Brazil nicolaaricciieri@yahoo.com.br

Abstract. Agent-Based Modelling and Simulation (ABMS) is a research methodology for studying complex systems that has been used with success in many social sciences. However, it has so far not been applied in education research. We describe some of the challenges for applying ABMS in the area of education, and discuss some of the potential benefits. We describe our proof-of-concept model that uses data collected from tablet-based classroom education to model how students interact with the content dependent on their engagement level and other properties, how we are calibrating that model, and how it can be validated.

1 Introduction

Agent-Based Modelling and Simulation (ABMS) is a methodology for studying complex systems, and is employed in diverse fields such as economics [26], social sciences [8], biology [12] and transport engineering [3]. [16] describe ABMS as being “particularly suitable for the analysis of complex adaptive systems and emergent phenomena”. Despite pedagogy studying what can clearly be described as a complex system [5,14,15], ABMS has, to our knowledge, not been applied in studying education.

There are a number of possible reasons for this, but in our opinion the most prominent one is that to apply ABMS, one needs a lot of available data, and a sufficient understanding of intermediate constructs and how they can be extracted from the data. This type of research is performed in the twin fields of Learning Analytics and Educational Data Mining, which have only recently reached sufficient maturity for ABMS to be a valid approach in researching various aspects of education. In this work, we aim to promote ABMS as a methodology for pedagogical research; in particular when studying classroom environments. We will give an overview of the state-of-the-art in ABMS, and describe some ways in which it could be applied to education research. Moreover, we will present a proof-of-concept ABMS for modelling a classroom education environment, and how we can use it in both the research and practice of education.
2 Background

Agent-Based Modelling and Simulation (ABMS) is a computational research methodology to study complex phenomena. In particular, an agent in this methodology is an autonomous (computational) individual that acts within an environment. The aim is (usually) to model the agents and their interactions, and study the resulting emergent properties of the system. It has long been known that very simple systems can portray complex behaviour. Probably the most well-known example of this is Conway’s game of life [9], in which a cellular automaton with four simple rules can, dependent on the starting scenario, have such complex behaviour that people are still discovering new patterns.

ABMS can be seen as the natural extension of such cellular automata, in that it considers the system as being distributed: rather than the system having a single state, each entity (agent) has a state, and agents decide individually (and asynchronously) on their next action, rather than the system as a whole moving from state to state. Nevertheless, they have in common that both ABMS and cellular automata are bottom-up (aka micro) models, where the behaviour of the system emerges from the behaviour of individuals, as opposed to macro models, which aim to (usually mathematically) describe the system as a whole without worrying about individuals. Because ABMS allows for distinct types of agents within the system, it is rapidly gaining popularity in biology and social science research, where describing the entire system in a set of mathematical equations is often too complex to be of much use, if possible at all, whereas it is possible to build sufficiently realistic models of individuals and how they interact with each other, and study properties of the system that emerge from the behaviour of many individuals.

For the development of an ABM, it is necessary to be able to calibrate and validate it [4]. The former refers to the process of tuning the model to best correspond to real data. In particular, this requires that there is detailed data on individuals’ behaviour in order to calibrate the agent models, data on their interactions, in order to calibrate the interaction models, and finally, data on the emergent properties of the system in order to validate the model. The validation of an ABM, or any model for that matter, is a complicated issue and there is ongoing research on how best to approach this [13,20], but without doubt, it is only possible with sufficient real data against which to validate.

Up until very recently, such data was not available for the field of education. However, over the last decade, many areas of education have been making increasing use of technology: (i) on the administrative side to keep track of students, their grades, attendance rates, etc.; (ii) at the course level, where LMS applications help teachers distribute course material, collect assignments and perform many of the day-to-day tasks of administering a course; and (iii) at the class level, mostly in the form of MOOCs, where teachers provide a digital course for students to follow worldwide (usually completely separately from the normal curriculum). All such applications generate, and store, detailed data at the level they are designed for. An administration system will store a history of students and classes throughout time, whereas an LMS tracks students through
a course, in combination with the material each student accessed, and created, throughout the course. Finally, MOOCs and other systems within a class collect detailed usage data for each student, logging each action a user takes throughout the class. The use of such data to perform data-driven analysis of pedagogical measures is a development of the last few years.

In particular, Learning Analytics is defined as: “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.” —Siemens [25]. Learning Analytics and Educational Data Mining are both relatively young, but fast-growing areas of research that focus on finding algorithmic methods for quantifying various metrics that can model a student’s learning process. There is a large overlap in the two areas, both in goals and in methodologies, with the differences between the areas being mostly in the philosophical and historical roots of the fields [24]. However, as Gašević et al. [10] point out, in accordance with our own findings, “there has been a dearth of empirical studies that have sought to evaluate the impact and transferability of this initial work across domains and context”, and “instructors expressed their preferences of learning analytics features that offer insights into learning processes and identify student gaps in understanding over simple performance measures”. Eberle et al. [7] discuss 12 grand challenge problems for Technology-Enhanced Learning. Among these challenges, we distinguish the similar themes: they identify challenges on how to model students using continuous temporal data, and how to use the data in providing appropriate feedback to empower teachers. Consolidating: while there is a significant amount of work [6,21,22,23] into descriptive and predictive measures, what is missing is how different measures can be combined into a theoretically sound, data-driven model of student learning.

We propose that Agent-Based Modelling for Education can address this question.

3 ABMS for Education

The integration of cognitive theory and agent-based approaches in technology-enhanced learning has so far been the particular domain of Intelligent Tutoring Systems (ITS) [11]. In this approach, the focus is on individual learning, and the student is helped by an artificial tutor. Cognitive theories are employed to model user knowledge, and a diverse range of AI techniques, including intelligent agents and recommender systems are used to decide how the tutor should act, what learning object to present to the user, etc. An example of a modern ITS approach is SimStudent [19], which uses machine learning techniques to update a rule-based agent system to adapt to students’ needs. A similar approach is presented by López Bedoya et al. [18], which uses a combination of computational techniques to plan, and replan coursework based on heterogeneous characteristics of the students. Such systems serve a different purpose from what we propose: rather than attempting to model students as socio-cognitive systems in order
to understand them, these systems model students as parameters for adaptive computational tutors.

**Fig. 1.** Overview of Components of a Digital Teaching Platform with intelligence models

Figure 1 gives the outline of components in a contemporary system for technology-enhanced learning and outlines what kind of models can help improve our understanding at each stage of education. We see that at the individual level, the modelling follows a cognitive modelling approach [1], and the aim, as best as possible, is to model the psychology of learning, and understand such concepts as creativity, learning styles, engagement and performance. This usually treats students as individuals in isolation, relying for instance on data from MOOCs where students can mostly be considered to be working through a course alone; interacting only with the content and not (much) with their fellow students or the educator. A system comprised of multiple agents like this could model a classroom. In this model, in addition to individual properties, the social interactions come into play. We distinguish between implicit and explicit social interactions. Implicit interactions are those that arise from individual actions that predicate a response. For instance, the teacher changes pages, and expects the students to follow suit. Explicit social interactions are those actions that are directly aimed at a social goal, such as discussing an item, asking a question, solving an exercise together. MOOCs and LMS systems attempt to capture many such interactions, although it is often impossible to do so: if interactions happen outside the system, the data is usually not available. The emergent properties
from such a system can also be studied at two levels: firstly, the interactions can have an effect on individual students, and secondly we can study properties of the classroom, both as aggregates over the individuals, like average performance or engagement, and properties that only arise at the level of a classroom, such as sociability. Finally, we can study how such systems are situated in the environment; the environment is generally described as everything that is modelled, but is not an agent. In the education scenario this is the course material and the possibilities for students and educator to interact with it, but also external parameters, such as noise level, light level, temperature, time of day and day of week, etc. With regards to the content, the data that is most easily manipulated by a computational model is the metadata, which describes properties of the actual course material, such as whether it is textual, a video, or interactive content, what subject(s) the content is related to, and for what level it is intended. By modelling the content, and other aspects of the environment, one can answer other types of questions, such as whether the same exact class at a different time of day would have led to better overall performance? Or what effect switching textual content for video content might have. This is the use of ABMS we have started to study, and we present our proof-of-concept class simulator in the next Section. However, we can think of other potential uses of ABMS for education.

ABMS can also be applied at a higher level of abstraction, in order to assist in policy decisions about educational institutions, teaching methodologies and content. Students might be modeled as having to decide what institute to study at, taking into account the policies and teaching philosophies. Or the institutes themselves can be modeled as agents. The modelling of such macro-level phenomena requires more administrative and demographic data.

A different, but complementary, approach for using ABMS in education is to use the ABMS as an educational tool for modelling other systems (e.g. a biological ecosystem). While that is not the use of ABMS we are interested in this work, it may nevertheless be insightful to give students control over an ABMS simulating their own situation, and allow them to experiment with what types of interventions might improve their own, their class’s or their school’s performance.

4 Proof-of-concept Model of a Classroom

We built the proof-of-concept Classroom Simulator to model a classroom in which the students are following the teacher during a lecture. It involves the simulation of the students’ interactions based on different educational contents and teacher characteristics, however it does not take explicit social interactions into account. In simulating a classroom, it takes three different steps: (I) Initialization, (II) Generate Signals, (III) Content Analysis and optimization. This aims to simulate the actual data that is output by a tablet-based digital teaching platform that we developed [17] for use in classrooms. This system generates detailed usage logs for each student, and the teacher. The idea of the model is to be able to generate such usage logs artificially, and simulate some specific types of
interventions. In particular, we aim at modelling teacher-oriented interventions, such as changing the type of material or teaching style. The simulator focuses on a subset of the log data, specifically those related to navigation through the content.

Step (I), generates the variation of students signals’ based on the characteristic of the teacher, the content and a class profile. Input in this step are the duration of a class, the educational content profile, the students’ profile and the teacher profile. The educational content profile is the number of different contents, their difficulty, their type (image, video, text or assessment) and their order. The teacher profile is related to this content: how orderly he goes through the material and what types of content he spends more time on. Students’ profiles are generated stochastically, by drawing random samples from a class profile. The class profile defines the number of students, and normal distributions for engagement (the amount of time a student spends working on the same content as the teacher), activity (the number of actions per timestep), and orderliness (whether they navigate through the content linearly from start to finish, or a haphazardly fashion that jumps between contents and navigates backwards) of the students. Furthermore, we simulate eye tracking time series data about whether an individual student is looking at, or away from the tablet at every timestamp. All of these parameters can also be extracted from the real usage logs generated by our tablet-based system. In addition to these profiles, a set of rules are generated (also adding random noise in the generation process) based roughly on the possible values each of the different parameters might take for each student, and the states these create within the system, the next state for any student. These rules describe for instance that if the student has the profile of being highly engaged, medium active, highly orderly, is looking towards the tablet, and has been studying the first content, which is difficult, and of type video, for at least 10 timesteps. Moreover, the teacher is looking at content 2, of type image, then the student will change pages to content 2. Given the possibility for combinatorial explosion, these rules are generated randomly without overlap until a sufficient (parameter of the model) number have been generated.

In Step (II), the class is simulated. All students, and the teacher, are initialized on page 0. For each timestep, the teacher and each student is evaluated: for each agent the list of rules are run through, and a distance from the agent’s state to the state described in the precondition of the rule is computed. The rule with the minimum distance is selected, and if this distance is greater than a threshold, the rule is executed and the agent transitions to a new state. The output of the simulation is a log of the events that occur during the simulation run.

In Step (III), we can calibrate (and also validate) the model by analysing the simulated log with regards to student engagement, orderliness and activity, and see if these correspond to the student profiles generated in Step (I). The calibration can be done by many different optimization algorithms, but in particular, we believe Genetic Algorithms are appropriate, with as fitness criterion a distance measure between the generated student profiles, and the input student
profiles. We start by generating different class, teacher and material profiles and generate a population of rule sets. The rules are evaluated through simulation, and then a subsequent population of rules is generated. The best rule set is selected. This rule set can then be tested with real data: compute the engagement, orderliness and activity level of a real classroom based on the logs from a session of use of the tablet-based education platform. Then use the rule set to generate an artificial log (which could also be compared to the real log for additional validation), and based on the artificial log, compute the different profiles again. We can then see how far they differ.

The final step is a work in progress, and we are iterating the genetic algorithm approach to calibrate our simulator.

5 Discussion

Agent-Based Modelling and Simulation is a powerful tool for doing research into complex systems, such as education, and while it has so far not been used in this field, we believe the data is now available to take advantage of this methodology. Moreover, data-driven research into education is a fast-growing field of research, and ABMS can take advantage of increasingly sophisticated quantitative metrics of education to incorporate into the agent models. Once agent-based models are sufficiently robust, they can even help validate these metrics, by quantifying how much explanatory power they add in the model, in comparison to the added complexity. This can be an invariable tool in quantitative research into education, by helping to hone in on what metrics best explain and predict student learning, and present such information to the students, teachers, parents and other stakeholders. Moreover, simulations can be designed to test what type of interventions work best.

Our own agent-based model, although still in an early stage, is intended in such a capacity. The model as described thus far is a first step, and we have specific extensions planned to better take advantage of the unique possibilities of agent-based approaches: the rules that are created to differentiate between agents’ behaviour do not take social interactions into account, and the student profile does not take motivational factors or learning styles into account. Moreover, we are still working on creating a valid model based on the data we gather, but the potential is clear: we can take real class profiles and test how they perform using a different content profile. If the students are more engaged than their input profile suggested, this is an indication that perhaps that type of content works better with that class. We can design experiments to test such hypotheses and use a methodology like design-based research [2] to iteratively improve the model and test its predictions.

Nevertheless, a lot of work still remains to be done, and if anything, this work should be seen as a call to arms. The ground conditions are there for using ABMS to both study, and improve our methods of education!
References

Using Semantic Web Technologies to Describe an Educational Domain

Tiago Thompsen Primo¹, Helder Coelho², Rosa Viccari³, and André Behr³

¹ SRBR - Samsung Research Institute Brazil, Campinas, SP, Brazil  
tiago.t@samsung.com
² Universidade de Lisboa, Departamento de Informática,  
Lisboa, Portugal
³ Instituto de Informática, Avenida Bento Gonçalves 9500,  
Porto Alegre, RS, Brazil

Abstract. This research presents a method to describe LO as Semantic Web Ontologies. The proposed method divides the Ontologies among three layers. The first is composed by the knowledge domain, the second by the objects that are part of this domain and the third are used for reasoning. As a practical example, it is presented the description of a subset of the LOM and OBAA metadata standard as an Ontology for the Layer One, also, an ontology of a LO for the Layer Two, and an Application Profile Ontology for the Layer Three. Those ontologies were all developed with OWL-DL. This work presented reasoning method was until now not explored by Semantic Web educational applications.

1 Introduction

Nowadays the availability of alternatives for the acquisition of information is extensive. The television, internet and mobile devices are platforms that make it easier to access various types of information in different situations. Considering such context, what if we could use the internet as a huge source of information for learning systems in a personalized matter?

One of the possible scenarios to illustrate that question regards to the fact that students and educators are users of such technologies, and in some situations they use a combination of devices to perform one task.

In the case of educators the reuse of part of a documentary for a class as complementary content, and in the case of students, YouTube videos and internet articles are shared with colleagues to perform a homework activity or with an educator as a content suggestion at Inverted Education [12] scenarios.

To allow such interoperability between different applications seems to be only a matter of protocol communication, but, if we consider each action as a signal that can be interpreted to infer students and educator characteristics, it becomes important to define not only the communication protocols, but also an interoperable knowledge representation.

Learning Object(LO) communities and other research groups are proposing standards since the early days of the Web to cope with such scenarios. What
changed so far is the amount of content, where and how to find it and how the content can be made available. Thus, the focus of this research is to integrate the standard technologies provided by the Semantic Web (SW) with Artificial Intelligence methods with the purpose to achieve personalized education in the context of the WEB as a huge repository.

In [16], the authors argue that educational content discovery should be personalized based on contextual factors such as prior knowledge, learning goals and learning pace. Our approach makes use of Linked Data [7] methods and technologies to represent domain knowledge allowing an expressive representation of user requirements and context. In addition, the educational content itself can be modeled in a way that facilitates its discovery and utilization not only by a specific repository, but in the context of the Web.

Given an expressive user model it is possible to focus at the prediction of user needs and more importantly the pedagogical needs. Hence, our goal is to understand and provide alternatives to identify the needs automatically, aiming to deal with the information overload in a proactive manner. To accomplish this goal, this work proposes an alternative to model Learning Objects using ontologies.

The variables related to the learning process of a student can be perceived from numerous perspectives (e.g. Pedagogy, Philosophy, Psychology, Human Computer Interaction)[9]. It is necessary to use a flexible and extensible approach to model the student, the interactions with a system and their LOs. Therefore, we propose the utilization of the standard languages from the SW together with reasoning techniques from the Descriptive Logics.

In this paper we will present our ontological proposal to describe LO, Students Profiles and their relationships allowing a two-fold contribution: (I) An ontological knowledge representation to describe LO, Students Profiles and (II) a practice to describe axiom with description logic to perform basic inferences and to support Intelligent Agents.

This representation and reasoning scheme allows the user model to have multiple facets, such as context, level of instruction and cognitive learning style. Those facets can be also modeled as an Agent.

The core idea is that the user agent is composed by a set of annotations that are defined and maintained by specific ontologies. Besides modeling the user, it is possible to adopt the same approach to represent the LO. In this paper, we restrict the content to Learning Objects [8], allowing us to open space for Agents to deal with incomplete metadata information and different metadata standards.

Related to the objective the contributions are: 1 - A method to describe LO as Ontology individual compatible with the SW; 2 - A method to describe the relationships among LO and 3 - An ontology engineering approach to describe ontologies and axioms to reason over this specific domain information.

To present our contribution, we start by Related Work. Following we present our Knowledge Representation for the Educational Domain, followed by our current results, discussion and our conclusions and future directions.
2 Related Work

The research of [6], presents a comparative study among several other works that made use of ontologies and SW to support educational applications. Such alternatives can be classified in three generations. The First generation is characterized by the Learning Management Systems (LMS), whose example is Moodle. The second generation is characterized by the use of AI techniques to support the learning process, for instance, the Intelligent Tutoring Systems and the third is characterized by the use of ontologies and the SW, an example is the current proposal that intend to cope with educational systems in a decentralized manner. Although they relate the importance of utilizing SW and ontologies, they do not provide a general methodology to engineer those ontologies, their relationships and a real case usage scenario.

Following this third generation of educational systems, as proposed by Dicheva, [1] presents two approaches using ontologies and SW techniques to support instructional design activities, in other words, allowing personalization alternatives to the pedagogical engineering process. Along with this, they present the concept of anytime-anywhere learning based on SW techniques. Their synergy with the present work lies in the use of SW and ontologies for reasoning. Although their proposal is mainly based on SW techniques, they do not provide alternatives to fulfill the communication gap between this alternative and the current learning systems.

The work of [10] presents an initiative related to this. They reinforce the need of broad solutions to cope with large-scale educational personalization. It means that the current learning systems and technologies must not be discontinued immediately and they need alternatives to migrate to a semantic environment.

In the work of [15] they are concerned with the data that is generated by the use of educational applications. In their proposal they mention that this data should be available by other applications and to the users. Using the SW is an alternative to this, specially when dealing with description and linkage of educational applications. The underlying idea, is that currently we already have several educational applications and we have to recommend the suitable application according to a student need. Albeit they acknowledge that SW can be used along with legacy learning systems, they have not proposed any alternative to describe and share the users characteristics and usage history.

This approach is also based on the research of [14], where one of the areas that need to be explored by educational systems, is the ability to have intelligent software that can help learners to filter information for quality and importance. The Semantic Web is a rich source to accomplish this [4], along with current interoperability issues that must be considered to cope with this challenge. Over the next sections we will describe our approach to represent educational content and user profiles using domain ontologies.
3 Building the Knowledge Representation

Users in the educational domain can be perceived accordingly to several perspectives (e.g. pedagogy, psychology, cognitive sciences and social behavior). In addition, a user profile is dynamic and often is composed during a learning process, instead of being previously determined. Therefore, our computational user model is specified using domain ontologies based on descriptive logic. This representation design allows us to cope with incomplete information and also to manage the consistency of the profile during updates. More importantly, the grounding on descriptive logics provides inference services to automatically discover implicit information using classification (subsumption) [13].

For the purpose of this work, we will focus on the use of the OWL due to its compatibility with the SW, its ability to provide decidable reasoning mechanisms and its status as a W3C recommendation.

The proposed ontologies are divided in three layers. The Layer One is composed by ontologies that describe metadata schemas, for instance, the LOM Ontology; the Layer Two ontologies, describes a User Profile, LO or their relationships with properties from a Layer One Ontology; and the Layer Three Ontologies comprehend the description of Applications Profiles that will provide reasoning over the Layer One and Layer Two.

To describe the method several key terms were used, that are technical terms of OWL ontologies. They are described as follows: 

**Class**: Describe concepts of a domain, structure that can encompass a set of Data Properties or Object Properties and individuals; 

**Properties**: Is a binary relation on individuals; 

**Object Properties**: Relations between individuals; 

**Data Properties**: Relationships among individuals and a XML Schema Datatype value or a literal; 

**Axiom**: A premise or a point to begin the reasoning process; 

**Range**: Links a property to either a class description or a data range; 

**Domain**: It is used to link a property to a class description; 

**Cardinality**: It is a form of restriction, defines the maximum or minimum number of individuals to link with a property; 

**Individuals**: They represent the objects in the domain that we are interested in.

3.1 Describing Layer One Ontologies

The Layer One ontologies are used to describe classes and properties that are used to represent individuals in the Layer Two ontologies. This layer stores ontologies with the semantics of some metadata schema, their cardinality, ranges, properties and axioms that are necessary to describe any application domain.

An application domain can be considered: A Standard for User Profile representation; an educational metadata standard; a relationship standard; among other possible top layer descriptions. For instance, Layer One ontologies com-
prehends the LOM Metadata Standard\textsuperscript{4}, the OBAA\textsuperscript{5} metadata standard, the FOAF\textsuperscript{6} metadata standard.

To design Layer One ontologies, we propose the following set of practices: Metadata became Class and Subclass; Properties became DP and OP according to their semantic; The semantic of a metadata became restriction axioms.

### 3.2 Describing Layer Two Ontologies

The Layer Two ontologies has the purpose to describe a User Profile, a Learning Object or a usage relation between them. Those ontologies are characterized to have import relationships with a Layer One ontologies and to be composed by individuals with data properties and Object Properties to describe them.

Usually we can apply some reasoning mechanism to verify the consistency of an individual through some Layer One ontology. For instance, if we describe a LO with the OBAA Standard, we can verify if the cardinality, range and value space were properly used. Also, due to the characteristics of an Ontology, if some description is incorrect we can apply an explanation algorithm to help a user to correct his/hers mistake.

Considering storage purposes, those ontologies can be stored in some formal repository, e.g. a Triple-Store, or even simply defining an URI for its access. This alternative gives flexibility to content designers that can simply build and publish their contents freely on the Web.

To engineer the Layer Two ontologies we proposed the following practices: A User Profile, a LO, a usage context or a relationship, each one of them represented by one ontology, and create as much individuals as necessary, with DP or OP to describe it.

### 3.3 Describing Layer Three Ontologies

The Layer Three ontologies are mainly developed to represent Application Profiles (AP). [11] defined an AP as compositions of metadata elements from one or more metadata schemas to provide a reduced set of elements that still retain compatibility with the original schema. For this work we extend this definition mentioning that an AP is an abstraction that stands for the practice to derive some knowledge based on axioms applied over this “ontology network”. This reasoning process, can be used to derive, for example, the users that have some pedagogical aspects, if a community is intended for persons with special needs, among other.

Basically an AP ontology will be composed, at least, by a class with an axiom that will infer the individuals that has their characteristics. For instance, it can

\textsuperscript{4} The Learning Object Metadata Standard [8] has an element set used to describe LO, it is considered a reference for such activity.

\textsuperscript{5} The OBAA metadata standard [17] was build mainly to extend LOM with platform interoperability aspects. It is a Brazilian Standard.

\textsuperscript{6} The Friend Of A Friend vocabulary is used to describe the relationships among Agents for the Semantic Web.
be created a class UsersWithSpecialNeeds with an axiom that describes that the property hasVisual must be present in an individual to be inferred as an instance of this class.

If necessary OPs can be added to improve some reasoning capability, for instance, an AP that has to many classes. Also, it is possible to relate different AP ontologies to compose one single ontology to increase the inference capability. These alternatives can lead to performance issues due to computational aspects.

To engineer a Layer Three ontologies we create as many classes with domain axioms as necessary to derive knowledge.

4 Sample Application

In order to present our proposal we will describe a Learning Object made use of the three ontology layers presented on section 3. For that, we will present the Educational Metadata Standard for the Layer One ontology, the LO ontology in the Layer Two and a Application Profile ontology in the Layer Three.

4.1 The Layer One Ontology

The first thing to consider while designing a Layer One Ontology, is related to the fact that they will provide all the properties that are fundamental to describe the Layer Two and three ontologies. Besides that we need to provide an URI for each ontology.

To design a Layer One Ontology, we will use as a case study the LOM Metadata Standard for three reasons: LOM is considered an international de Facto standard to describe LO’s [5]; it is common to the educational technology researchers and there is no LOM OWL ontology available to reuse that cope with the current proposal.

LOM is an extensive educational standard, we will present the description of a few classes and properties that can be extrapolated to build other ontologies. The LOM ontology is available to reuse through the following URI 7.

Considering the limited space, we shall focus on the LOM LifeCycle group due to the fact that it is relatively small but preserves the semantic complexity of the larger groups, such as General or Educational. Following we present the Classes and Properties described with the previously proposed ontology engineering approach.

The choice of the LOM LifeCycle relates to the fact that this group is small but preserves the semantic complexity of the larger groups, such as General or Educational. Following we present the Classes and Properties described with the previously proposed ontology engineering approach.

Classes Each metadata from LifeCycle group became a class, and a subclass with cardinality restrictions according the standard. For instance, Contribute is a subclass of Life Cycle and has a cardinality max 30.

7 http://gia.inf.ufrgs.br/ontologies/LOM.owl
Properties Properties can be classified as Data Properties (DP) or Object Properties (OP). DPs are the data itself. OPs describes the relationship between classes and individuals. OPs are also associated with the metadata cardinality.

Those cardinality restrictions, when applied to OP, divide and group individuals that are distinct among each other, but have a relationship with another higher individual. For instance, take as example the Figure 1.

In the Figure 1, at its center, it is illustrated a sample LO Individual that is divided between two parts. The number one (1), illustrates a generic LO representation model; The number two (2) illustrates the usage of the OP Contribute, in this case, named LOM:hasContribute. As can be seen there are three individuals represented. The higher Layer LO Individual and two other individuals linked by the OP LOM:hasContribute and each one of them, with specific DP.

This kind of relationship allows, for instance, the reuse of the individuals LO + LOM:hasContribute + ID1 and/or LO + LOM:hasContribute + ID1 in different versions of a LO.

This example were prepared to exemplify the description of a LO with such ontology engineering method, the next section will present the characteristics of the Layer Two Ontologies.

4.2 The Layer Two Ontology

The Layer Two Ontology intends to provide an ontology that represents LO, User Profiles, Educational Applications, among other that can make use of a Layer One Ontologies. An LO is composed by an ontology and a set of individuals described by properties that are imported from a Layer One Ontology.
For instance, the description of a LO with the LOM ontology properties is made by: Convert or create a LO; create an OWL file to represent the LO; import the LOM OWL ontology; describe the individuals that represent the LO; create OP relationships when necessary.

To describe our sample LO, we shall mention the LO Ramis. This LO was developed with two metadata standards, IEEE LOM and OBAA. The underlying purpose for its creation was to provide an interoperable object that is compatible with tree hardware platforms: Internet, Digital Television and Mobile Devices.

Ramis has a set of metadata that primarily was not intended to be represented by proposed method. With that in mind, we start our conversion by analyzing those meta-informations and illustrate its complexity with Figure 2.

![Diagram of Ramis OWL file](image)

**Fig. 2.** The representation of individuals for the Ramis OWL file

Figure 2, has the indication 1 that emphasizes the higher layer individual; the indication 2 the OP hasRequirement; the indication 3 the OP hasPlatformSpecificFeatures; the indication 4 emphasizes the OP hasSpecificRequirement; The indication 5 the OP hasSpecificOrComposite and finally the indication 6 the OP hasOrComposite. Each individual has its own set of Data Properties.

### 4.3 The Layer Three Ontology

The Layer Three ontology provides reasoning over the Layer Two ontologies. To this proposal this type of ontology is classified as Application Profiles (AP). An AP, is an ontology composed by a class or a set of classes that describes specific domain knowledge and has at least one axiom to infer which individuals, from the Layer Two ontologies, are members of those classes.
Reasoning is mainly useful to verify if an LO is properly described according to some standard, to provide inferences according to specific characteristics, infer new relationships between LO, to support some statistical process, among others.

To describe an OWL AP ontology we must consider: What is the knowledge to be derived from the Layer Two ontologies? Is it possible to retrieve it with a SPARQL query? Reasoners are not extremely powerful, one axiom is enough? How many classes are necessary?

As an example, we shall describe an AP that will infer a LO, only if this LO has a specific set of metadata. Considering this, we present the OBAA Lite AP. This AP, was developed by [3]. Mainly her research explored the OBAA and LOM standards to identify the minimum set of metadata necessary to describe a LO. The Figure 3 presents the class and the axiom built to infer which LO can be inferred by this AP.

![OBAA Lite Ontology Axiom](image)

Fig. 3. OBAA Lite Ontology Axiom

The current version of *Ramis* individual ontology does not get inferred by the OBAA Lite application profile because it does not comply with its axiom. In order to have it properly inferred, we had to add the missing DPs and OPs to cope with the OBAA Lite AP. The modification has resulted in the classification of *Ramis* as member of the OBAA Lite AP class as can be seen in Figure 4.

5 Discussion

The work of [6] states that educational systems based on the web can be classified among three generations: The first, examine centralized architectures and are based in proprietary model of educational contents; the second is centralized and supported by AI techniques, they adopt personalization based on user profiles, but also maintain the proprietary model of educational contents; the third is not centralized and based on Agents and service oriented architectures, supported by ontologies coherent to the principles of the SW. The state-of-the-art educational applications are between the first and second generation. This fact motivates the
appearance of initiatives that stimulates the transition to the third generation. This work is an example initiative.

Also, [2] divides the SW educational applications in the three columns. This work was able to contribute with the two first columns. To cope with the first column, an alternative to reduce computational costs related to explore SPARQL queries for simple application profiles. E.g Search LO with a specific property value. To cope with the second column, our major challenge will be to reuse the educational ontologies that are already available and are not compliant with SW, also privacy is a delicate matter, specially in this case when dealing with private and personal information.

A related research opportunity is to conserve and securely migrate the information of LO and User Profiles among those generations. Such effort may lead to a better user profiling and the singularity of educational technologies.

An alternative to this, the usage of service-oriented architectures emerges as an alternative for the transition among generations, their major characteristic is to focus on the definition of a standard communication format between the applications in the web. Using such approach leads the communication between different applications in distinct programming languages and/or architectures.

Beyond communicating among educational applications, to understand what is stored is crucial to provide personalization alternatives. Provide computational structures that can manipulate the information and derive knowledge, autonomously, became fundamental. The tendency to achieve such interoperability is explored by the use of ontologies. The presented approach provides a method to describe and store information allowing extensions, connection between information and compatibility across multiple domains.

---

8 A W3C Recommendations to query RDF files, it can be accessed in http://www.w3.org/TR/rdf-sparql-query/
To cope with this tendency, it had opened space for the creative development of service-based applications mainly supported by AI techniques. Those applications can derive knowledge, provide inferences, retrieve content from different servers, build flexible networks to support educational applications, stimulate the relationships between students and teachers that are geographically distributed and mostly to provide the possibility that the users being immerse in a permanent learning environment.

6 Conclusion

There can be several application development alternatives for the educational domain. We can explore the authorship of LO, development of Agents for personalized student courses, suggestion of educational materials among others. The present work had presented an alternative to represent User Profiles and LO by ontologies to allow reasoning a reasoning alternative for educational applications.

The presented ontologies made use of the three layer proposal to describe the knowledge domain. In each one of them we described and presented an example of them. The LOM ontology is considered complex due to its specifications and characteristics. This fact, in some cases, caused the reasoner to be overwhelmed with ontologies of the Layer Two that where composed by many individuals. An alternative could be, separate each group of the LOM metadata in a single ontology. In the best case scenario the reasoner would only have to cope with a limited set of properties.

The Ramis ontology made use of properties from the LOM and OBAAL ontologies. This Layer Two ontology had to deal with a complex set of DP and OP to be described. Although it can be considered a complex ontology, in an automatic process we could obtain several interesting benefits such as: Easy update, for instance an individual that is updated can be linked through an OP; the reuse of some individuals by other LO, for instance an technical individual that is common to several other LO; the described individuals can be validated according to a Layer one ontology; the storage can be made by a URI; it is possible to build relationships between LO ontologies by properties and it is compatible with the current SW stack.

The method to build a Layer Two ontology can be used to describe the user profile ontologies, educational domain ontologies, relationship ontologies, or any other that might describe an educational activity. Such amount of relationships can lead to performance issues, specially in the reasoner.

There is a lot of work that yet need to be done, specially when considering students privacy matters, content usage rights, security polices over such information and, not most important, public polices that stimulate and popularize the principals of open knowledge. This can lead to an large scale evaluation that can measure the effective of this approach for the current learning system.

As future work we will explore a Triple-Store alternative to index and store Layer Two Ontologies such as LOs, User Profiles and Relationships between them, a Service Oriented alternative to integrate this proposal with some current
educational application and describe new APs, to evaluate if the use of OWL-DL is the most adequate solution to represent such kind of knowledge and explore the automatic conversion of Legacy LO repositories according to this proposal.

References

Workshop 3

First International Workshop on Intelligent Mentoring Systems

(IMS 2016)

https://imsworkshop.wordpress.com/
Preface

1st International Workshop on Intelligent Mentoring Systems

In the world of professional development and lifelong learning, mentoring has become an important ingredient to assist learners to progress, to transit from one phase into another (e.g. from formal education to on-the-job training). Within digital learning environments, it is timely to provide a new breed of intelligent learning systems that provide mentor-like features to promote learner’s ‘self-actualisation’. Crucial for intelligent mentors will be the ability to help learners connect their learning that is usually acquired through digital resources with the real world, which brings forth the key challenge that the workshop aims to address.

Mentors who are to facilitate self-actualisation require a broad (but may be shallow) understanding of the learner, and the current situation in order to select appropriate pedagogical strategies and respond in a motivational, emotionally-aware way. Accordingly, Intelligent Mentoring Systems (IMS) require multi-faceted learner experience modelling mechanisms to get sufficient understanding of the learner, his/her current situation, and relevance to past experiences by the same learner (or by other people). IMS will also need strategies for appropriate interaction with the learners, and to promote reflection and forward planning.

This workshop will make a major step in consolidating effort and shaping a research community by providing a forum that explores the following questions:

- What are the underpinning theories, models, and technologies for Intelligent Mentoring Systems?
- How will Intelligent Mentoring Systems support learners and tutors?
- In which domains/contexts do we expect to see Intelligent Mentoring Systems?
- What are the key research aspects to address next?

The workshop will include two main sessions of paper presentations, each followed by a group discussion.

The first session includes position papers discussing key underpinning for intelligent mentoring systems. In ‘Some Challenges for AIED Systems in Taking on Long Term Mentoring’ Du Boulay sets out key challenges faced by designers of systems for mentoring learners over longer periods of time. In ‘Intelligent Mentoring Systems for Making Meaning from Work Experience’ Dimitrova et al. present a vision for intelligent personal assistants to foster meaning making from work experience. In ‘Personalised and Adaptive Mentoring in Medical Education – the myPAL project’ Van Labeke et al. argue that design-based research approach is required for designing personalised adaptive mentors. In ‘An Active Learning Model Employing Flipped Learn-
“Gamification Strategies” Dicheva and Dichev propose a novel instructional approach, called gamified mentored learning, which combines elements of flipped learning and course gamification.

The second session presents experimental studies that inform the design of intelligent mentoring systems. In ‘Dynamics of Trust in Group Peer Mentorship’ Adewoyin et al. present a study of group peer mentorship which found that peers’ rating behavior is influenced by their trust score and it depends on the roles that they assume in the mentoring relationship. In ‘Group Learning, Student Clustering and Peer Mentorship’ Pepeseu et al. present features for adaptive group learning and peer mentorship. In ‘Usability of an Active Video Watching System for Soft Skills Training’ Lau et al. report initial results from an experimental study with an active video watching system that provides self-regulated learning scaffolding for soft skills learning.

Each paper was reviewed by two members of the workshop program committee.

Workshop Program Committee:
- Vincent Aleven, Carnegie Mellon University, USA
- Roger Azevedo, North Carolina State University, USA
- Ben Du Boulay, University of Sussex, UK
- Paul Brna, The University of Edinburgh, UK
- Darina Dicheva, Winston-Salem State University, USA
- Alexandra Poulouvassilis, Birkbeck College, University of London, UK
- Nicholas Van Labeke, University of Leeds, UK
- Phil Winne, Simon Fraser University, Canada

Workshop Chairs:
- Amali Weerasinghe, University of Adelaide, Australia
- Vania Dimitrova, University of Leeds, UK
- Lydia Lau, University of Leeds, UK
- Antonija Mitrovic, University of Canterbury, New Zealand
Some Challenges for AIED Systems in Taking on Long Term Mentoring

Benedict du Boulay

Dept. of Informatics, University of Sussex, Brighton, BN 1 9QJ
B.du-Boulay@sussex.ac.uk

Abstract: Many AIED systems have been developed that help learners in terms of their motivation over the period of a single session or problem-solving episode. This paper sets out some of the challenges facing designers of AIED systems for mentoring learners over much longer periods. These include (i) understanding the learner’s value system and their reasons for engaging in the learning; (ii) the ability to introduce new material and conclude material for a single session in the context of the overall learning journey; (iii) the ability to track and react to the learner’s motivational trajectory over a much longer period than simply short term frustration (say) with a particular issue; (iv) the ability to help the learner keep track of their progress over a long period – including their progress in learning how to learn and learning how to self-motivate.

Keywords: Mentoring, learning, motivation, meta-cognition, meta-affect, meta-motivation.

1 Introduction

Historic work on motivation, as applied in Artificial Intelligence in Education (AIED) systems, was based on a model that tried to detect the learner’s evolving motivational state and then react to it in a motivationally sensible manner, either by adjusting the learner’s tasks, the feedback on the learner’s work, or the manner in which the tasks were offered [1-3]. More recent work in this area within AIED has expanded its repertoire of emotions associated with learning [4] and developed methods both to detect emotions [5] and behaviours symptomatic of poor motivational states, such as gaming the system [6], boredom, frustration, confusion and engaged concentration [7], as well as to react to them in sensible ways [8].
Despite these advances, the temporal focus continues to be largely on short term events during a single lesson or episode of problem-solving, with little concern for how that lesson or episode was initially pitched or introduced in the first place [9], how it might be best summarised at its conclusion, or how it fits into the longer term issues associated with the whole course or module, or indeed with the learner’s long term educational goals. In many ways these kinds of issue are regarded as the responsibility of the human teacher and not the AIED system that has only temporary charge of a little bit of learning.

2 Longer Term Mentoring

Designing AIED systems for longer term mentoring raises a number of extra issues beyond those already being tackled by current AIED systems. Rather than simply reacting to the learner’s motivational state, systems for mentoring need to be more proactive in the way that they help the learner deal with new learning experiences and reflect on past experiences. The need for this can be seen, for example, in the concern about the drop-out rate of students making use of MOOCs as a consequence of their largely un-mentored form of learning [10].

First is the issue of the greater investment of time and energy of the learner in seeing a whole course or longer period of study through to a successful conclusion. More care is needed at the start for discussing with the learner about whether what is proposed is the right course of for them. This immediately brings in the issue of the learner’s goals and values, as well as their expectations (whether accurate or not) of what the outcome will be and what the process will be like [11]. Understanding the learner’s value system and why they want to engage in a long-term piece of learning is crucial to mentoring. One only needs to consider the discussions one has with prospective PhD students to know that they can be profoundly mistaken about what they will achieve, how their experience will unfold and how they will feel at various stages.
Second, while AIED systems currently concerned with motivation try to deal with frustration and confusion in the short term, these issues are magnified in the context of longer periods of study. We have all faced setbacks to our creativity – the blank sheet of paper syndrome or loss of self-confidence in the face of lack of progress. Over the long term becoming demotivated can gel into something more difficult to deal with – losing heart with respect to the value of one’s long term goal or of one’s ability to reach it.

Just as expert human teachers include among their goals “first, to sustain and enhance their students’ motivation and interest in learning, ... and second, to maintain their pupils’ feelings of self-esteem and self-efficacy, even in the face of difficult or impossible problems” [12], so PhD supervisors will often spend considerable effort helping their students manage their “bumpy” motivational trajectory over three or more years. However, two advantages of working over longer time-scales are that one does not always need to react immediately to issues and can advise a student to take time out from a difficult task in order to regain emotional and motivational equilibrium.

There are predictable issues normally associated with learning a new skill or concept; for example that performance in the short term can be impaired while the new skill or concept is gradually integrated successfully with what was already known or achievable. The degree to which this kind of setback is managed successfully depends in part on the learner’s own theory of learning [13]. Over the longer term it can become harder to believe that repeated mistakes are productive and that determination and the willingness to keep trying will always win out. Here the mentor can help the learner to reflect on their own experience, drawing their attention to past successes in similar circumstances, for example.

My own recent experience in learning Argentine Tango gave me some sharp insight into the frustrations associated with learning a brand new skill over a long period. Progress seemed absurdly slow. Steps that the teacher executed with ease were hard even to describe to myself, let alone mimic, yet in many ways they ap-
peared so simple. Steps that had been learned initially one way had to be relearned when new factors needed to be taken into account, such as holding my own body more upright or holding my partner more closely. And, of course, the mistakes were all too public: “I’m a professor, why am I finding it so difficult just to put one foot elegantly in front another”. An important role taken by my teacher was precisely the mentoring role of helping me appreciate that I really had made progress when I thought I had not.

Third is the issue of the learner losing track of where they are in the whole process, developing tunnel vision, and neither being able to see the overall goal nor the role of the current activity in achieving it. One role for the mentor here is to occasionally take the longer view and help the learner appreciate what has already been achieved and what the best next step is. This issue of introducing and concluding new activities in ways that indicate the direction to be travelled and the distance already covered takes on greater importance as learning timescales lengthen.

3 Meta-motivation and Learning how to Learn

Longer term periods of learning often have aims that go beyond the material of the course, and these may well need to be explicitly mentored. First is the issue of learning how to learn, and indeed learning to enjoy learning [14]. One of the potentially valuable outcomes of the setbacks and disappointments mentioned in the previous section is that they form a bedrock of experience that the learner can, with help, exploit. For example, they can come to understand better how to engage in complex learning activities over a long timescale, what the likely problems and setbacks are that they may need to face, and how they can develop personal strategies that work for them in overcoming these disruptions.

Related to the above is the issue of meta-motivation, i.e. one’s own understanding of one’s own motivational processes. Meta-motivation involves both meta-cognition and meta-affect [15]. Those who acquire both the insight to understand their own inner motivational lives and the ability to regulate their degree of moti-
vation equip themselves well for most other learning journeys. Here the role for mentor is to help the learner reflect on what works for them in terms of strategies for self-motivation.

4 Motivational challenges

The main motivational challenge raised by working over the longer term is that the mentor needs to understand rather more about the goals and values of the learner, about their theory of learning and about their meta-motivational capability. Just pointing a camera at their face, instrumenting their heart-rate or skin resistance are not going to yield the sort of information needed to help the learner at difficult points in their learning. While to some extent one can use general exhortations to try harder or offer praise for achievements, these are just not going to be sufficient for the learner who has lost their way, and possibly also lost heart, in the value of what they are doing. This is a hard issue and probably depends on the ability to have some kind of interaction with the learner about why they are engaged in the learning and how they feel about their progress [16]. A second challenge is logging both the work done by the learner and their motivational trajectory over the long term, so that feedback to the learner can refer to earlier points in the learning: “Remember when you thought you could not solve that problem, but in fact you found a way – well I think we are at a similar point”. This is easier to deal with in that it extends existing logging mechanisms. A third challenge for a long term mentor is to be able to anticipate difficulties for the learner before they occur and help the learner prepare for them. This needs different kinds of domain and student model whose focus is as much on learning process issues – fear of the blank sheet of paper – as on domain level outcomes. This would need tasks and potential learning experiences to be categorised in terms of their motivational consequences for different types of learner.

5 Conclusions

This paper has set out some of the challenges in building tutors that extend learning from a single session or a short sequence of
interactive sessions to a longer term mentoring role. These include (i) understanding the learner’s value system and their reasons for engaging with the learning; (ii) the ability to introduce new material and conclude material for a single session in the context of the overall learning journey; (iii) the ability to track and react to the learner’s motivational trajectory over a much longer period than simply short term frustration (say) with a particular issue; (iv) the ability to help the learner keep track of their progress over a long period – including their progress in learning how to learn and learning how to self-motivate.

The implicit model of mentoring indicated above is a development of ELM-ART [17], a system for teaching programming that referred the learner back, when facing a difficulty, to a similar previous problem. Some small steps in this direction in terms of referring the learner back to previous metacognitive and motivational states as well as to previous problem-solving have already been undertaken in the area of learning programming [18].

6 References

5. Arroyo, I., et al., *Emotion Sensors Go to School*, in *Artificial Intelligence in Education: Building Learning Systems that Care: from Knowledge Representation to Affective Modelling,*


Intelligent Mentoring Systems for Making Meaning from Work Experience

Vania Dimitrova\textsuperscript{1}, Alex Poulavassili\textsuperscript{2}, Nicolas van Labeke\textsuperscript{1}, Tamsin Treasure-Jones\textsuperscript{1}, Miriam Zukas\textsuperscript{2}, Paul Brna\textsuperscript{1}

\textsuperscript{1}University of Leeds, United Kingdom
\textsuperscript{2}Birkbeck College, University of London, United Kingdom

Abstract. This position paper presents a forward-looking view on addressing a long standing professional learning challenge faced by higher educational institutions, namely assisting students to make meaning from work-based experience and develop as reflexive professionals. We suggest that a synergetic approach, building on existing research in professional lifelong learning and intelligent learning environments and taking advantage of new opportunities provided by emerging technologies, will underpin a new breed of intelligent mentoring systems for professional learning. They will foster the learners’ meaning making process, as well as assist tutors in their roles as coaches/mentors.

Keywords: Professional lifelong learning, reflexive practitioners, work-based learning, intelligent learning environments, intelligent mentoring systems

1 Motivation

It is vital that higher education meets the growing demands for professionals who are able to adapt to dynamically changing contexts, respond to new sociocultural challenges, and play leadership roles within complex heterogeneous systems. Education has to prepare reflexive practitioners who are lifelong learners capable of creating, transferring and modifying knowledge across different settings. One of the most effective ways to meet these aims is to include work-based activities within subject-based education [1]; indeed, this is seen as a key priority for the development of better-skilled workforce and boosting economic growth and social prosperity [2]. Educational institutions are progressively introducing work-based practical experience, ranging from fully embedded practice throughout the whole curriculum, to short-term industrial experience in the form of placements and internships, to problem-based learning drawing on realistic situations.

However, there remains a longstanding and seemingly intractable challenge, namely connecting formal subject-based education and informal work-based learning. This manifests as the need to support re-contextualisation of knowledge across different learning settings [3]. Reflexive practitioners explore their experiences to better understand their practice and improve themselves and the work environment. Unlike reflection, which is perceived as a somewhat passive instrument of observation, reflexivity
takes the form of a meaning-making activity in which the learner considers assumptions, makes interpretations, identifies perspectives, draws connections, and decides on actions [4]. This is in essence a creative process which involves the construction of personal narratives that represent the individual’s perceptions of their work experiences. Meaning making is insufficiently supported in higher education settings:

- Learners produce shallow, often inaccurate or superficial, reflections on their work-based experiences; they overlook key points, fail to recognise crucial cues, cannot make comprehensive interpretations, are unable to see all perspectives, and make poor connections across activities.
- Learners do not make the best use of their work-based opportunities; they do not prepare adequately, fail to ask appropriate questions, do not seek out experiences which will help them develop, do not solicit timely help from those they work with, and they fail to use tutor support appropriately.
- Tutors are unable to provide effective coaching/mentorship support: from the limited data available, they may have a very partial understanding of the learners’ experiences, limitations or inadequacies across various work activities; they are unable to facilitate learners to be proactive and engage with their work practice.
- Workplaces may have to undertake remedial education with new employees: although new workers always require training, employees who are inadequately prepared through their education will need substantial training, coaching and support if they are to succeed in their transition into work.

Technology can play a key role in innovating education practice in general, and reflexive learning through work-based experience in particular, by providing scalable and cost-effective interventions. We present here a forward-looking view of the role of intelligent technology in augmenting reflexive learning by supporting learners to make meaning from their work-based practice and assisting tutors to act as coaches/mentors facilitating the linking between work-based experience and formal education. The paper outlines a vision for developing intelligent mentoring systems for professional learning (IMS-PL). It draws on research in reflexive learning which provides the pedagogic underpinning, and points at requirements for the role of technology (Section 2). We then briefly review existing strands in intelligent learning environments that can provide the starting point for IMS-PL (Section 3). Finally, we list key challenges and suggest possible ways to address them when shaping IMS-PL.

2 Pedagogic Underpinning

The pedagogic underpinning for IMS-PL is based on contemporary approaches for reflexive learning and professional development, which identify key aspects of making meaning from work experience.

The embedding of work-based learning activities within formal education provides stimulating means for deeper learning, allowing learners to apply not only their subject knowledge and skills but also their interpersonal and intrapersonal skills [5]. Transitions from formal education to the work-place, or between different work-place settings, can be framed as critically intensive learning periods which challenge the
learner/practitioner to quickly acquire new knowledge specific to the new setting, to understand new organisational structures and cultures, and to establish new working relationships [6].

Most commonly, education focuses on reflection, i.e. learning and developing through examining what we think happened on any occasion, and how we think others perceived the event and us. Reflexivity is potentially more complex than being reflective: reflexivity requires understanding the myriad ways in which one’s own presence and perspective influence knowledge and actions [7]. In other words, reflexivity is a stance of being able to locate oneself in the picture, and to appreciate how one’s own self influences actions. Reflexivity means “finding a way to stand outside of ourselves to get a more objective view of ourselves” [7]. Reflexive learners are those who are able to contextualise and re-contextualise their knowledge and skills when moving between different settings of formal, workplace and informal learning. They engage actively with the development of their own knowledge and skills, and also with the social and organisational context in which their learning or workplace practice is taking place.

Reflexivity is associated with meaning making from work practice where the learner assesses, makes deeper interpretations, explores connections and alternatives, and decides on actions [4]. Meaning making is rooted in work practice that is inherently contextualised (it exists in a particular setting, and in a particular social and organisational context) and embodied (practitioners carry it out in both mind and body) [8]. Thus, aspects such as emotion, motivation and self-awareness cannot easily be separated from cognitive aspects, and reflection on practice will necessarily comprise subjective, as well as objective elements. Engaging in ‘learning conversations’ helps students to articulate and critique their growing understanding of practice [9]. In order to capture this dialogical dimension of learning, “confrontation either by self or others must also occur” [9].

Confrontation by self can occur through the creation of narratives, which allow learners to link their current experiences with previous experiences in an inherently personal way, enabling them to critique both themselves and their environment [10]. Bolton proposes the use of writing for exploring our own actions and those of others around us, and argues for narrative writing as a fundamental tool for meaning-making [7]. This enables the learner to see not only ‘in-the-mirror’ but ‘through-the-mirror’, exploring different perspectives and setting a course of action. Collier brings to the fore the place of creativity in the reflection process, arguing that the imaginative aspect of reflection allows movement away from a “tick-box” mentality [11].

Narratives facilitate meaning making from experience to interpret what has happened, make sense and reflect. This resonates with the vision of designing intelligent systems for inclusion, self-esteem, recognition of different perspectives, self-awareness, and reflexivity [12]. Narrative has been used, e.g. as theatre role play activity or as story writing activity, to foster creative thinking and make reflection more motivating and deeper [13]. However, recent studies show that today’s millennial learners are exposed to a vast amount of interactive media which imposes a radical change to the production and consumption of content [14]. Hence, new forms of narratives will be required which are closely aligned with digital media production.
The timing of reflexive activities is also very important as it can influence the quality and nature of the reflections: if reflection occurs immediately after an event of heightened emotions it is likely to be more subjective than if it occurs some time later, and thus a sequence of reflections over time is needed to draw out a deeper interpretation and understanding of the experience [15]. Guidance and supervision are key to reflective practice and are factors that learners perceive to be beneficial to their learning; and hence tutors have to be effective coaches and mentors [16].

Following these arguments, it can be expected that IMS-PL need to help learners to become aware of their emotions during work practice and can use emotion incidents as triggers of ‘confrontation with self’ in the form of personal digital narratives linking personal reflections across various work practice contexts. At the same time, acknowledging the key role of tutors in facilitating productive reflection, IMS-PL need to support them to gain a better understanding of their learners’ experiences and thus provide more engaging and meaningful coaching/mentoring.

3 Technological Underpinning

The technological underpinning for IMS-PL is provided by recent developments in intelligent environments for adult learning. We list below some of the current models and techniques which can be the basis for IMS-PL. Note that this is not an exhaustive list, we have only selected technological solutions that can be related to the requirements drawn in Section 2. For example, social systems and collaborative environments have been used for reflective learning from work-based practice (c.f. [17]) but the effective use of such environments implies that the learners already possess critical thinking and creative imagination to produce useful reflections. They have been used to support reflective learners who are already progressing in their professional career (e.g. doctors or public administration employees). The key challenge that inexperienced professionals, such as students at universities or vocational training colleges face – how to make meaning from their work-based activities – remains largely unaddressed and unsupported. Although collaborative environments can provide ‘confrontation by others’ to trigger reflection, they do not tackle the important problem of developing learners’ critical and creative thinking abilities so that they can more effectively remember and interpret their work experience, draw connections and associations, consider alternatives and link with professional development. It is the fostering of this intertwined thinking for augmenting reflexive learning that is the prime focus for the IMS-PL vision we shape here.

Intelligent learning environments are now a mature topic in research and practice. They build on the considerable body of research in Artificial Intelligence in Education (AIED) to create adaptive learning environments capable of adapting to learners and of providing personalised support. AIED researchers are starting to look outside the mainstream educational environments towards adult education and workplace learning, to provide innovative learning models that are universal, inclusive, lifelong and seamlessly integrated in everyday practice (c.f. [18]). As stressed in a recent review [19], in the next 10-15 years AIED will play a key role in supporting adult lifelong
learning. A possible way to do this is by providing ‘lifelong learning companions to advise, recommend, and track learning’ [19]. We expect that IMS-PL will include such companions and envisage that personalised virtual assistants for reflexive learning from work experience will be part of the next generation of AIED for adult learning.

IMS-PL will have to address cognitively complex domains which are associated with 21st century transferable skills, e.g. communication, collaboration, critical thinking, creativity, emotional intelligence, decision making. These skills are gaining growing attention in the AIED community, within the well-established stream of intelligent learning environments for ill-defined domains [20]. Example systems to support learning in ill-defined domains include collaborative spaces building on social interactions or situational simulations where the learners can ‘experience’ complex situations in a safe environment, e.g. military training [21] or simulated internships [22]. Critical for supporting learning in such open environments, where multiple interpretations and perspectives exist, is to provide effective learner support, adapted to the learner’s cognitive and affective states. We envisage that adaptation to both cognition and affect will be a key feature of IMS-PL.

Open learner modeling, where the learner is presented with a model that the system has obtained about them based on interactions monitored by the system, can foster reflection and meta-cognition [23]. Furthermore, when the open learner model is used as a trigger for interaction with the learner, i.e. interactive open learning modelling, the learner can be provided with reflective scaffolds to help learners to understand themselves [24]. We envisage that IMS-PL will leverage interactive open learner modeling substantially extending the interaction to provide mentor-like assistance to help the learners develop an understanding of their strengths and abilities, see and interpret different perspectives and opportunities, make sense of personal experiences, develop strategies and set goals, and improve their confidence and self-esteem.

In addition to the above, we expect that IMS-PL will offer new interaction spaces to promote reflection and self-awareness contextualised in practice, which will build on the blurring boundaries of the physical and digital worlds. This will leverage emerging technologies that enable the development of smart coaches empowering people to take control of their lifestyle (e.g. the quantified self), as outlined below.

4 A Vision for IMS-PL

Following the discussion in the previous sections, we hypothesise that IMS-PL will include a new generation of personalised virtual learning assistants which will be embodied and situated in real-world practice and will provide scalable, contextualised and longitudinal support for making meaning from work experience. They will leverage key emerging technologies, such as wearables and personalised mobile assistants. This technology ensemble has already brought disruptive innovation in other areas (e.g. quantified self) empowering people to become aware of their behavior and take control of their lives. We list below key research questions that have to be addressed to achieve this vision, and outline possible ways to tackle these questions.
**RQ1**: How to capture, in an unobtrusive way, the learner’s emotions and affect during his/her work practice and make the learner aware of the emotions he/she has experienced during work activities?

This can be addressed by using wearable sensors (e.g. on smart watches or bands) which can detect low-level signals and ‘emotion episodes’ with spikes of positive or negative emotions. This can be combined with emotion detection from speech/text (e.g. the learner taking notes on a smartwatch or mobile phone before, during, or immediately after a work activity). An emotion dashboard can be provided to the learner to make him/her aware of their emotions.

**RQ2**: How to trigger critical and creative thinking to promote reflection-on-action by utilising the sensed emotions during work experiences?

This can be addressed by extending existing note-taking tools with interactive contextualised ‘nudges’ to foster the learner’s critical and creative thinking. This will help the learner to recall relevant aspects of their experience, interpret it, link it with past experiences, and set goals for future actions. A nudge framework can be implemented in the form of short dialogue games, using the sensed emotions as triggers for interaction and utilising knowledge-enriched semantic analysis of the learner’s notes.

**RQ3**: How to help learners to create digital narratives, linking their current experiences with previous experiences and exploring alternative ways for interpreting work activities?

This can be addressed by offering an interactive story-telling environment where learners can inspect their experiences and ‘make a story’, i.e. a coherent narrative linking together different experience assets, such as notes and digital objects (e.g. pictures or presentations related to the experience). Support can be provided by offering story templates to create narratives linked to transferable skills and by personalised prompts directing the learner to aspects they may explore further.

**RQ4**: How to assist tutors in their new roles of coaches/mentors who facilitate meaning making from work experience?

This can be addressed by providing an experience exploration dashboard for tutors, so that they can inspect their learners’ reflexive narratives and become aware of key aspects related to their work experiences. Tutor support can include notification and visualisation tools based on semantic analysis and exploration of the learners’ experiences, reflections and narratives, enabling tutors to provide timely and meaningful feedback to their students.

Experimental studies will be required to assess to what extent the above features support reflexive learning and meaning making from work experience. These will focus on selected work-based activities, e.g. work placements or project internships, within specific domains where the technology has already been used in similar settings and where work experience is a crucial addition to formal education (e.g. Medicine or ICT). It would also be useful to trial any system in environments that are not technology rich, e.g. field-based construction work. This will provide constraints that
will bring sharper focus. Evaluation of our proposed pedagogical and technological framework will also require selecting and examining the development of specific transferable skills, e.g., communication, collaboration and innovation, as well as looking into the broader skills of learning to learn and professional development. Such evaluation can be undertaken using qualitative methods, engaging with all stakeholders in co-design and co-evaluation. In addition, quantitative methods (e.g., surveys) can be used to assess technology acceptance and perceived usefulness. Evaluation of the tutor support will need to similarly examine its benefits and possible drawbacks in assisting tutors in their roles as coaches and mentors.

**Individual differences** will play a major role, and additional evaluation studies will need to examine whether, and how, individual differences affect the effectiveness of IMS-PL. Possible factors to include are demographics (gender, age, nationality); cognition, emotion regulation, and personality traits; self-regulation and curiosity. These user characteristics can be combined with analysis of the learners’ and teachers’ interaction data with IMS-PL during user pilots. Based on this, strategies informing how to include personalisation and adaptation in IMS-PL can be derived.

**Acceptability** of the new technology will be a key driver for the new research stream. It is possible that the new technology impinges on the practitioner’s work practice and relationships with colleagues. This would raise various kinds of issues related to social acceptability and ethics – these would have to be identified at the outset and carefully monitored as research progresses.

5 **Concluding Remarks**

In this paper we have motivated and presented a forward-looking view on addressing a long-standing challenge faced by higher education institutions, namely assisting students to make meaning from their work-based experiences in order to develop as reflexive professionals. We have identified possible roles of intelligent technologies – including wearable sensors, personalised mobile assistants, interactive open learner modelling, contextualised nudging frameworks for creative critical thinking, interactive story-telling environments, and notification and visualisation tools for tutors – in innovating reflexive learning through work-based experience by providing scalable and cost-effective interventions for both students and tutors.

Our ultimate goal is to shape a new breed of intelligent learning environments, which we term *intelligent mentoring systems for professional learning (IMS-PL)*, which include mentor-like features to help learners become reflexive practitioners and establish themselves in their professional careers, while at the same time recognising the key role of tutors as coaches/mentors facilitating learners’ productive reflections and learning.
References

Personalised and Adaptive Mentoring in Medical Education – the myPAL project

Nicolas Van Labeke¹, Jane Kirby², Trudie E. Roberts¹

¹Leeds Institute of Medical Education, University of Leeds, Leeds, UK
²Leeds Institute of Health Sciences, University of Leeds, Leeds, UK
{n.vanlabeke,j.kirby,t.e.roberts}@leeds.ac.uk

Abstract. This position paper describes a long-term Technology-Enhanced Learning initiative at the Leeds Institute of Medical Education in which a personalised adaptive learning mentor will be deployed for all MBChB students enrolled in the course. The system, myPAL, is enriching the existing TEL programs embedded in the curriculum and will be leveraging recent advances in Learning Analytics and Open Learner Modelling. The paper presents the context of the project and the opportunities that deployment settings will offer, and highlights the research and development strands that will underpin it.

Keywords: Medical Education, workplace learning, adaptive learning, Design-based Research, intelligent mentoring system

1 Introduction

This position paper is a first attempt in describing a long-term Technology-Enhanced Learning research and development agenda that is being unrolled at the Leeds Institute of Medical Education (LIME¹): the design and development of a personal mentor – aptly called myPAL – for students in medical education. The aim of the project is to enrich an existing program of Technology in medical education (digital resources, computer-based assessments, mobile learning) by applying current advances in the field into our educational context, notably Open Learner Modelling and Learning Analytics [1, 2].

But as we will briefly illustrate, the specific requirements of medical education (e.g. practice-based learning) means that the focus myPAL will not so much on tutoring students through the curriculum – even in intelligent ways – but more toward mentoring them throughout their developmental pathway(s) across the educational and professional settings. And as the workshop is rightly querying about, questions will need to be addressed throughout the design and development of the system as to how we are supporting learners in that process (e.g. self-regulation, motivation), what are the tools and mechanisms (e.g. modelling, analytics, visualization, reasoning) that need to be

¹ http://medhealth.leeds.ac.uk/info800/leeds institute of medical education/
deployed, and what are the conditions for learners to adopt and appropriate such a tool in the long term (e.g. workplace learning, lifelong professional development).

2 Context

The context of the work on myPAL is the 5 year undergraduate course program leading to the degree of MBChB (Bachelor of Medicine and Bachelor of Surgery) which allows successful students to provisionally register with the General Medical Council and start supervised practice of medicine (a further Foundation Year program being required for unsupervised practice in the UK). The MBChB curriculum is a challenging program based on professional values and core themes that are integrated throughout the five years, in what is usually described as a “spiral” curriculum [3]. In this structured learning approach, students are introduced, during the first year of the degree, to the core principles and themes that underpin clinical practices and form the foundation on which later years will be coming to again and again, building on what students have already seen and done.

At the same time, students are increasingly moving away from the lecture theatres and traditional academic delivery of foundations into placements and clinical settings, their growing experience and ability allowing them to progress on an “entrustability” scale (from observe to supervise, initiate and then peer teach), expressing higher level of attainment (and responsibility) in clinical settings.

The Leeds Institute of Medical Education has the responsibility to design and deliver the MBChB curriculum for the University. One of the aspects of the innovative approach is the extensive adoption of Technology-Enhanced learning in the curriculum. For more than 10 years, the Technology in Medical Education (TIME) team has been developing and deploying digital resources to students, working closely with clinicians, academics, students, patients and carers to ensure quality and relevance. Students have been encouraged to use mobile technology, initially through the delivery of PDAs to every undergraduate students but increasingly through a Bring-Your-Own-Device paradigm.

What we are now considering is how to bring that experience even further by enabling a more personalized and adaptive learning experience for students in medical education.

3 myPAL – Personalised Adaptive Learning

So what does this context means for an innovation and research agenda in the context of Technology-Enhanced Learning and Medical Education?

---

2 See https://www.medicine.leeds.ac.uk/curriculum/ for a description of each of the 5 years of the MBChB at the University of Leeds
4 See the TIME website at https://time.leeds.ac.uk for a list and description of digital resources and computer-based assessment systems.
• The MBChB degree is recruiting about 250 students each year, which over a 5 years curriculum (plus intercalated year and Foundation Years) gives a cohort of significant size for data collection, longitudinal studies and volunteer-based co-design activities for our research and development projects.
• The “spiral” nature of the curriculum ensures that the learning focus and experience of the students is consistent and accumulated upon previous learning interactions, a feature that most modular, standalone academic undergraduate degrees don’t share.
• The presence of a Technology-Enhance Learning team embedded in LIME and including researchers and software developers gives us the flexibility to develop our own systems in-house, enabling a much tighter interaction between research and development, in particular with methodology such as Design-Based Research and Agile prototyping.
• The widespread usage of computer-based approaches for the curation and delivery of digital resources (e.g. app- or web-based eBooks, educational videos, revision applications) and for the handling of academic assessments (e.g. summative assessments, reflective or end-of-placement reports) already gives us a wide and diverse range of learning and interaction data that could be exploited.
• The early adoption of mobile technology by students means that a comprehensive set of data covering the whole 5 years of the curriculum by a complete cohort of students, from entry to graduation, is now available for performance and predictive analytics.
• The mix of academic settings (lectures, workshop) and workplace settings (placements, practice-based learning) is a fundamental approach of the MBChB curriculum, to the point that “practice, performance and learning are so interlinked they are inseparable and dependent on the specific setting.” [4]. This will create tensions with aspects of a learning mentor, especially when some if the data evidencing sources and performance of learning are not easily collectable or even identifiable.
• Even if placements are a significant part of the MBChB curriculum, one-to-one tutoring or mentoring activities remain relatively time and resource consuming, and therefore remain under-exploited. Even with the presence of tutors or senior clinical staff on site, opportunities for feedback remains limited due to time or task constraints, and to tutors and learners (in)ability to deliver and – respectively – identify feedback.

This short analysis of the context is strongly indicating that is a niche for the deployment of an appropriate technological approach to mentoring, to support medical students during their learning and practice, in traditional academic settings but also, and more important, in work-based settings. How to design, deploy and study the adoption of such a socio-cognitive system will be the objective of our research agenda on myPAL.

4 A Design-Based Research Approach

What we investigate with the development of myPAL to exploit these opportunities will be prioritized in the following weeks and months, in particular to meet some of the MBChB timeline requirements. We already know that the design space opening up with
our vision of a personalised adaptive mentor for medical students will need to embrace social, cognitive, emotional and organisational aspects of learning and human-technology interactions. In order to systematically explore and evaluate the design decisions that will inevitably have to be made, the myPAL system will be designed, tested, trialled and evaluated with students in both controlled and real-life settings, following a Design-based Research (DBR) approach [5].

In recent years, DBR has become a popular methodology for moving learning systems beyond the stage of research prototype. Its main advantage is a pragmatic agenda of producing better artefact by utilising theory while advancing theory through the design and usage of these artefacts. The tighter integration between practice and theory, between researchers, practitioners and users allows the design and evaluation of interventions (technology-enhanced in our case) that are aiming at changing educational practices. In particular, the holistic approach of DBR in integrating real-life settings supports (even promotes) the exploration of emerging properties out of the interactions between technology, people and social spaces. The main disadvantage of DBR however is the theoretical and methodological diversity that is at the core of its premises and the sparse documentation and replicability of the procedures used in previous works. Frameworks such as LATUX [6] or Socio-cognitive engineering [7] will be valuable in shaping our own approach of DBR and contributing to its methodological fundations.

Rapid prototyping and Agile methodology will be used to ensure that the system is developed incrementally and embedded in the Design-Based Research cycles, that fixes, improvements and new features originating from the co-design sessions are integrated seamlessly and evolutionarily into the live system, maximising the chance of a long-term appropriation by students and institutions.

The work on myPAL will be organized around 4 parallel but intertwined strands (see Fig. 1) running, in the first instance, over the next 5 years of the MBChB curriculum: co-design of the main system; research and development of targeted functionalities; exploitation of historical data; and development of the technological infrastructure.

![Fig. 1. The four strands of research, design and development of myPAL](image-url)
4.1 Co-design cohort

This is to be considered as the principal strand of our work on the myPAL system, the overall objective being the study of its design and development with the new cohort of students starting their MBchB in September 2016, and its eventual adoption and appropriation over their interaction with the curriculum. This objective will be supported by series of participatory design activities spread across all stages of development of the system; initial co-design of ideas and innovation, testing of early prototypes, evaluation of impact, adoption and appropriation of the system, longitudinal studies of cohort (e.g. attitudinal shift to data privacy and sharing). For example, we will be exploring whether a dashboard is an appropriate interface for accessing learning-related data [8], or if a more adaptive, feedback-oriented interface might be defined.

4.2 Targeted Research and Development

The second strand of the myPAL project is the targeting of specific issues of research and development that could not be scheduled in the natural progression inherent of the cohort-led co-design activities. One such situation is to guarantee, as stated in the aim of the project, that every student will have access to some functionalities of myPAL as appropriate as possible according to their own progression in the MBchB curriculum. We might therefore have to focus on early development of functionalities that are more appropriate for Year 3 or Year 5 students, in order to keep them in the loop. For example, a significant part of the assessment process are Objective Structured Clinical Examinations (OSCEs), a competency-based assessment methodology that is linked with performance objectives, mapped to curriculum outcomes and is increasingly used in healthcare education programs [9]. OSCEs are a very concentrated – and stress-generating – experience where students are assessed on specific clinical competencies and their performance checked. But the strong competency frameworks underpinning their design – and the fact that feedback at such granularity is still not given back to the students – make them a very good candidate for developing the backbone for an intelligent mentoring system.

4.3 Historical Data and Predictive Analytics

As mentioned earlier in the document, we have now access to a large set of historical data on which to perform deeper predictive learning analytics. The 5-year cover of the MBchB curriculum will give us opportunities to explore long-term learner modelling and, once combined with the data collected with live students, comparisons and baseline, whenever appropriate.

4.4 Technological Infrastructure

The final strand is the development of the technical infrastructure (i.e. front-end, back-end, data warehouse, etc.) according to our needs and requirements for interoperability
with existing systems or libraries. For example, we are considering the use of xAPI as the metadata specification mechanism for the learning events being stored in *myPAL*’s Learning Record Store and exploited by the system and its associated analytics engines. The specification have been developed over many years and has reached a degree of maturity sufficient for observing a number of projects adopting it – see for example [10]. But a point made by many adopters of the specification – e.g. [11] – is that so-called *recipes* (i.e. the mechanism advocated by xAPI developers to standardise the expression of learning experiences) is a key condition for long-term adoption by the community. Therefore, we believe that our work with *myPAL* could play an important role in developing, testing and validating recipes for learning experiences related to mentoring activities in the context of medical education.

5 **Toward an Intelligent Mentoring System with *myPAL***

At an early stage of the establishment of Learning Analytics as a research discipline on its own, a very important paper [12] identified 4 challenges facing the community: 1) Build strong connections with the learning sciences (e.g. how is learning taking place); 2) Develop methods of working with a wide range of databases in order to optimize learning environments (e.g. using Learning Analytics outside the confines of VLEs); 3) Focus on the perspectives of learners (e.g. extend criteria of learning success beyond grades, personalised visualisation); 4) Develop and apply a clear set of ethical guidelines (e.g. ownership and stewardship of data).

The design and development of *myPAL*, and its continuous deployment in real-life settings, with cohorts of students in the MBChB curriculum, open several perspectives for addressing many of these challenges. The workplace learning approach, with students in placements expected to observe, interact and learn from their experience, will provide a wide range of real-life settings where user-centric technological solutions will be trialed and deployed to supplement the generation of adaptive feedback, the multimodal collection of new learning data, the creation of nudges to trigger deeper learning.

In an application context for *myPAL* such as a medical education curriculum covering a whole 5-year of academic and professional development, traditional intelligent tutoring approaches, at the level of topic or problems, are neither realistic prospects (for one, there is no cognitive tutor that will do the job), nor desirable (intelligent simulation-based learning on specific aspects of the curriculum would be).

But an *intelligent mentoring system* that will support the learner in transitioning from academic to workplace learning by appropriate feedback is clearly a timely and pertinent approach. We believe that many aspects of the *myPAL* project could lead to significant contribution to such an endeavor: a focus on self-regulation of learning through appropriate feedback on learning; adding social machines functionalities [13] to complement inadequate (or missing) semantic information; the design of appropriate feedback mechanisms, both implicit and explicit, that will operate seamlessly and timely in settings when the immediacy of self-reflection and action for changes will vary a lot.

---

5 xAPI, or Experience API, https://github.com/adlnet/xAPI-Spec/blob/master/xAPI.md
Advances in Open Learner Modelling and Learning Analytics will provide us with many of the concepts, tools and directions required to explore many of the issues that the notion of Intelligent Mentoring Systems are raising and that the workshop will undoubtedly elicit.

6 References

An Active Learning Model Employing Flipped Learning and Gamification Strategies

Darina Dicheva and Christo Dichev
Winston Salem State University
Winston Salem, NC, USA
{dichevd, dichevc@wssu.edu}

Abstract. The presented here instructional model is based on two premises derived from psychological and learning theories. The first is that students take greater control on their own learning, which implies active learning where the instructor acts as a mentor. The second is that the learning environment enables interplay between the extrinsic forces acting on learners and the intrinsic motives and needs inherent in human nature. We claim that such a model can be built by leveraging strategies from both flipped learning and gamification.

Keywords: Active Learning, Flipped Learning, Gamification, Motivation

1 Introduction

We are witnessing a change in the traditional teacher-centered instructional model, which has been established to fit well to the brick-and-mortar schools. That model had been affected by the boundaries of the traditional school – the physical, temporal, organizational, and other boundaries. For example, those boundaries have helped in managing the access to scarce printed resources; courses have emerged naturally out of the boundaries of classes, what teachers know, and the need for scheduling; terms are designed to fit the ‘natural’ boundaries of holidays, etc. [1]. All these boundaries target to solve the problem of making an efficient use of scarce resources.

Human learning in the before-school times has been boundary-free and significantly different. The natural learning has always been ubiquitous, contextual, activity-based, social, and led by intrinsic motivation. Subsequently, with the continuing advances in the technological arena and in the different areas of the society and the dramatic progress and increase of human wellbeing, it is just natural for some of the existing boundaries of the formal education model to be crossed. It is difficult to predict what exactly the future education will look like without all those boundaries, including the accreditation. Yet, various novel instructional approaches and strategies have started to emerge. While to a large extend they are more or less within the traditional boundaries, they do focus more on the learner and less on the standard lecturing. They feature more flexibility, ubiquitous learning, active learning, etc. Most of them rely on the use of online content and on self-learning. Examples comprise online courses, including MOOCs, and flipped classrooms.
A common problem of the two mentioned instructional models, which rely on learners’ self-study at home, is that learners often do not do their assigned work. This entails employing methods that can boost learners’ motivation and engagement so as to improve their performance and achievements. In this paper we propose a novel instructional approach that combines elements of flipped learning and course gamification: *gamified mentored learning*. Before presenting it, we briefly discuss flipped learning and gamification whose basic principles back the soundness of the model.

2 Flipped Learning

Flipped Learning is defined in [2] as “a pedagogical approach in which direct instruction moves from the group learning space to the individual learning space, and the resulting group space is transformed into a dynamic, interactive learning environment where the educator guides students as they apply concepts and engage creatively in the subject matter.” This model leaves class time open for interactive learning activities, which cannot or is difficult to be automated. In essence, “flipping the classroom” means that students gain first exposure to new material outside of class, usually via reading or lecture videos, and then use class time to do the harder work of assimilating that knowledge, through practicing, problem solving, discussion, or debates. In terms of Bloom’s revised taxonomy (2001), this means that students are doing the lower levels of cognitive work outside of class, and focusing on the higher forms of cognitive work (application, analysis, synthesis, and/or evaluation) in class, where they have the support of their peers and instructor [3].

The theoretical foundations that justify dropping the in-class lecture delivery include student-centered learning theories and methods such as active learning, problem-based learning, experiential learning, collaborative learning, cooperative learning, peer-assisted learning, and peer tutoring (see for example [4]). The student-centered learning theories provide the theoretical basis for the design of the in-class activities in flipped learning. Constructivism and collaborative learning stem from Piaget’s theory of cognitive conflict [5] and cooperative learning stems from the zone of proximal development theory of Vygotsky [6].

The flipped learning model has attracted significant attention in the educational community. Around 2012 there was a boom of publications on instructor and student perceptions on flipped classrooms considering it as a highly successful practice. Then there was a decrease of the enthusiasm with publications analyzing the pros and cons of teaching in a flipped classroom. However, recently a revival is noticed with teachers blending the flipped classroom approach into their traditional curriculum, using the method more selectively [7]. The overview of empirical studies related to the flipped classroom by Bishop and Verleger [4], concludes that students’ performance has been reported to improve compared to performance of students in a traditional classroom setting and student opinions tended to be positive, but there were invariably a few students who strongly disliked the change. Among the other findings were that many instructors instituted a required pre-class quiz on the lecture material; students
preferred live in-person lectures to video lectures, but also liked interactive class time more than in-person lectures, and shorter videos were preferred.

The flipped learning model is considered to be useful if students are motivated to do independent work and enjoy more collaborative in-class sessions [7]. This observation points to a serious challenge since as Sappington et al. [8] among others show, college students don’t generally complete reading assignments.

3 Gamification

A key premise for the success of the growing alternatives of traditional education, e.g. online and flipped learning, is students’ motivation and engagement in the instructional process. It derives from the need of more self-regulation, intrinsic motivation, time management, and independence of the learner. This implies that the students not only need to have the internal push to complete a task, but also to be able to complete the task independently and to keep themselves on track without constant monitoring.

On the other hand, games are well known stimuli that drive people to take voluntary actions in a predictable way. Thus a natural idea is to harness the characteristics of games that give rise to this phenomenon and put them to use in learning situations where engagement is lacking. Researchers have been attempting to isolate and identify the attributes of video games that stimulate motivation, engagement, and perseverance. This research has led to the “gamification” trend.

Gamification is the use of game thinking and game mechanics in non-game contexts to engage users in solving problems [9]. It has become a popular tactic to encourage specific behaviors and increase motivation and engagement. Though commonly found in marketing strategies, it is now being implemented in educational programs as well to help educators find the balance between achieving their objectives and catering to evolving student needs [10]. A number of instructors have been exploring the concept of gamification with the intention to use it as a tool for engagement and motivation. A systematic mapping study of the use of gamification in education is presented in [11].

The growing popularity of gamification comes from its potential to foster motivation, behavioral changes, friendly competition and collaboration in different contexts. The theoretical foundations of this are several motivational theories and models that can impact users’ behavior [12]. Maslow [13] explains that human behavior is need-based and goal-oriented: it is driven by people’s desire to satisfy physical and psychological needs and these needs are what motivate them into actions. Maslow’s hierarchy of needs is represented in a hierarchical pyramid with five levels: physiological, safety, belonging, esteem, and self-actualization. The four lower levels are considered physiological needs, while the top level is considered growth needs. The lower level needs must be satisfied before higher-order needs can influence behavior. Pink [14] hypothesizes that in the modern society where the lower levels of the Maslow’s hierarchy are more or less satisfied, people become more motivated by other intrinsic motivators. These intrinsic motivators are: autonomy, mastery and purpose which focus on our innate need to direct our own lives (autonomy), to learn and create new
things (mastery), and to do better for ourselves and our world (purpose). His work is based on the Self-Determination Theory (SDT) proposed by Deci and Ryan [15], which posits that humans continually and actively seek challenges and new experiences to develop and master. Self-Determination Theory states that to have a positive well-being, people need to feel that they have control over their situation, feel competent, and connected to others. Conditions supporting the individual’s experience of autonomy, competence and relatedness are argued to foster the highest quality forms of motivation and engagement for activities, including enhanced performance, persistence, and creativity. According to [16], in order to be intrinsically motivated to perform a task, a person must be kept in a state between an anxiety and boredom known as flow. Clear goals, a sense of control, immediate feedback and a balance between skill and challenge are some of the factors that contribute to flow.

4 Gamified Mentored Active Learning

The presented here instructional model is based on two premises rooted in the above described psychological and learning theories. The first is that the students have to have much more control on their own learning (promoting autonomy). This implies active learning where the instructor is acting as a mentor to the learners. The second is that the learning environment has to facilitate the interplay between the extrinsic forces acting on learners and the intrinsic motives and needs inherent in human nature (SDT). We claim that such a model can be built by leveraging strategies from both flipped learning and gamification. From the flipped learning we take the pre-class home reading/video watching and the in-class activity-based work. However, the latter features a far less controlling role of the instructor in class:

- Students complete pre-class work to familiarize themselves with the factual knowledge (e.g. by watching short videos and/or reading text).
- In the beginning of the class, the instructor answers questions on the pre-class reading and may present a short explanation for topics considered difficult.
- The active learning in class includes problem solving, collaborative work on projects, labs, discussions, group work, etc., mentored by the instructor.

Differently from the typical flipped classroom practices, no graded quizzes (including such with clickers) are recommended in class to test whether students have done the pre-class reading or in-class work, since this is a very strong demonstration of the controlling role of the teacher. Instead, we recommend promoting interest based on active involvement and variety, curiosity and challenges; fostering an environment where it is safe for students to fail, but which does not allow the failure to define them; breaking the in-class work into manageable steps coupled with instant feedback with optional grading for completion. One motivator supporting this approach is the availability of a variety of (automatically checked) practice exercises and quizzes for student self-learning and self-assessment, which completion should be additionally stimulated by relevant gamification mechanisms. This should also replace the typical for the flipped classroom practice of assigning after-class homework that usually leads to a very high load on students demanded by the flipped class.
Considering the gamification elements that can be used for gamifying a course, it is neither possible nor desirable to suggest specific game mechanics and dynamics since those depend on the specific course, instructor, students, context, etc. However, based on empirical evidences we believe that the following game mechanisms provide a good assortment for practical course gamification: accruing points, rapid feedback, freedom to fail, unlocking content, virtual currency, skill points, progress bar, leaderboard, avatars, rewards/incentives, and social engagement (collaboration and friendly competition). The guiding strategy for choosing a particular configuration of game mechanics includes: reward behaviors that are under students' control; reward effort not talent; create little quick wins at regular intervals; enable measuring progress and achievement; focus on students' individual progress rather than on their performance in relation to their peers; enable combining game elements with intrinsic factors.

We piloted the proposed model in a Data Structures course, aiming at reducing the high rate of drops and failures. We redesigned the course employing methods and techniques from both the flipped learning model and gamification. Flipping the classroom allowed us to introduce programming labs in class, which is not typical for the standard way of teaching this course. Unfortunately, we could not include everything that we wanted, since we did not have appropriate technological support, especially with regard to applying game mechanics and dynamics. The employed gamification elements included: accruing points (max 1000 points), rapid feedback (max 24 hours), freedom to fail (allowed multiple submissions of labs and home assignments), and social engagement (collaborative problem solving and pair programming). We wanted to include also unlocking content (early personalized release of content and labs), virtual currency (rules for earning and spending course bucks), skill points, progress bar, avatars, rewards, and a leaderboard, but we did not have appropriate support in the course delivery system used on campus (Blackboard).

Our experience of piloting the proposed instructional model revealed the necessity of a new type of educational software that can support intelligent mentoring of gamified flipped learning formal classes or informal groups of learners.

5 Conclusion

It is reasonable to assume that in the foreseeable future the formal education, which features accreditation, will continue to exist. While this implies that some boundaries will remain in place, we anticipate that many will fade away, in particular those that are related to the dominant controlling role of the teacher in class. The idea of having a mentor instead of a teacher is not new. The problem is that the instructor cannot be replaced by a mentor in the traditional educational process. Instead, new instructional models are needed that from one side ensure the learner-centered, active learning and from another, reinforce the intrinsic motivation for learning. In this paper we proposed one such method where gamification is used to effectively complement and support the flipped model of learning. As most instructional approaches, this model is not one-size-fits-all. Rather, it needs to be tailored to individual classes, students, and learning objectives. In this context we rely on two important assumptions for its suc-
cess. We presume that: (1) gamification is carefully designed and implemented by experienced instructors to suit the specifics of the learning context, and (2) an appropriate intelligent educational software support is available. Regarding the former, more studies are needed to generate empirical evidence for the efficacy of using different game mechanisms and their combinations in different learning contexts to improve student motivation, engagement, and academic performance. As to the latter, easy-to-use tools for automatic generation, checking, and personalized delivery of abundance of practice exercises are needed, as well as course gamification platforms that can efficiently support instructors of academic courses or mentors of self-organized learning groups who want to gamify learning experiences.

6 References

Dynamics of Trust in Group Peer Mentorship

Oluwabunmi, Adewoyin\textsuperscript{1}; Fred, Phillips\textsuperscript{2} and Julita, Vassileva\textsuperscript{3}.

\texttt{bunmi.adewoyin@usask.ca, phillips@edwards.usask.ca, jiv@cs.usask.ca}

Department of Computer Science\textsuperscript{1,3}; Edwards School of Business\textsuperscript{2}
University of Saskatchewan, Canada

Abstract. Group peer mentorship is a collaborative learning venture where peers are both mentors and mentees. Existing work had shown that trust is vital in building a strong mentoring relationship. In this research, we implemented a modified peer review process with a group of professional accountants, to support them in group peer mentorship. Our goal was to find out how peers’ interpersonal trust scores affect their rating behavior both as mentors and as mentees. Our results show that the interpersonal trust score influences rating behavior and it depends on the roles assumed by peers in the mentoring relationship.

Keywords: Interpersonal trust, mentorship, collaborative learning

1 Introduction

Group peer mentorship, according to [19], is an ideal form of collaborative venture where peers are both mentors and mentees. It is believed that peer mentoring is cost effective and that feedback from peers is more welcome (or at least less intimidating) than the feedback from an instructor [2]. Existing work on mentorship [10] discussed different factors that can influence the success of a mentoring relationship: communication, differing expectations between the mentor and mentee, appreciation of circumstances that affect each party, and trust. This study focuses on the effect of trust on peers’ behavior in group peer mentorship. Other authors have studied factors that influence the development of trust in a relationship; for example, [15] discussed three factors – ability, benevolence, and integrity; and [14] proposed some hypotheses based on these factors. However, we are not aware of any existing work that explores the effect of interpersonal trust on peers’ rating behavior in group peer mentorship. This is the problem addressed in this work.

We implemented a peer review framework proposed originally by us in [1] using an available peer-review tool, called PeerCevity [21], to help a group of students in a Master of Professional Accounting program improve their audit and analysis skills. The main objective was to see if there is a correlation between each mentor-mentee’s interpersonal trust score and the ratings the mentor and mentee gave to each other after each mentorship session. Thus, the research question in this paper is:
• “How do peers’ interpersonal trust scores affect their rating behavior both as mentors and as mentees?”

The rest of this paper is organized as follows. Section 2 presents related work in the areas of mentorship, peer review and interpersonal trust; section 3 describes the study method. Section 4 discusses the results of our study, and section 5 concludes the paper.

2 Related work

Mentorship is a relationship between a more knowledgeable person and a less knowledgeable person for the purpose of career or psychosocial development [23]. In traditional mentorship, it is believed that mentors have to be older and more experienced than their mentees. However, this is costly because it depends on the availability of the experienced mentors. One solution to this problem is to engage participants in peer mentorship, which brings together peers, close in age / career achievement. Research had shown that peer mentorship is the strongest source of students’ cognitive and affective development [4]. Peer mentorship is not a new concept and it has been used in workplace environments as well as academic environments where knowledgeable students mentor less knowledgeable ones, e.g. [9]. Group peer mentorship brings peers together for support and learning in groups, to achieve their academic, career or psychosocial goals [23]. This model of mentorship is believed to save time and cost, but its success relies on factors, such as the group cohesion and reciprocity [1], and trust in each other’s ability to perform [11]. There are different types of trust described in the literature [27, 34]. In this study, we focus on interpersonal trust.

Mayer, Davis and Schoorman (1995) defined interpersonal trust as the “willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party.” There are three parts to this definition of trust – willingness to be vulnerable, expectation and risk-taking because of the uncertainty of the ability to monitor or control [17]. With this definition, it is easy to think that people that are high in interpersonal trust can be gullible. However, research had debunked this assumed relationship between trust and gullibility [26, 35]. [26] described trust as “believing others in the absence of clear-cut reasons to disbelieve”. Also, [35] discovered that trusting people are more vigilant and detailed on the information they receive and process about other people’s trustworthiness. Therefore, they have the tendency to have a high expectation of other people’s trustworthiness.

In this study, online group peer mentorship is implemented with the use of peer review system, which is an example of an online collaborative tool. In scientific community, peer-reviewed articles are considered as trusted scientific communications that have gone through the quality control process of peer review [11]. Peer review has been trusted to deliver quality and novel research articles since its inception in the 18th century [31]. Although research had shown that peer review is susceptible to issues like bias, incorrect reviews that may arise due to misinterpretation of
the authors’ intentions and inconsistencies [31, 28], some improvements to the peer review process had been proposed [33, 32, 13, 1]. For example, double blind review minimizes the chance for bias and thus supports better the summative evaluation purpose (selection of quality papers) [29]. To support its formative objective, which is essential to support group peer mentorship, back-evaluation of reviews was proposed by [13, 1]. With back-evaluation, authors are required to evaluate their reviewers. Our previous work had shown that back-evaluation of reviews is preferred by peers and has encouraged them to give thorough and helpful feedback [2]. Also, in the peer review study by [3], it was confirmed that with back-evaluation, peers were not reciprocating their review ratings, but were being honest and helpful in their feedback to their peer reviewers; although peers used pseudonyms to mask their identities in this study. However, no work has been done to investigate the relationship between peers’ interpersonal trust and their ratings tendency, both as mentors and mentees. [15] defined three factors that contribute to the development of trust: ability, benevolence and integrity. Ability is defined as the set of skills that give a party influence in a given domain, benevolence refers to the trustee’s propensity to do good, help or show positive attitude towards their trustor, while integrity refers to the perception of the trustor that the trustee would keep to certain acceptable principles [15]. Based on this model of trust building by [15], [14] proposed some hypotheses to link interpersonal trust of mentees and mentors to ability, benevolence and integrity. We extend their work by exploring how interpersonal trust influences peers ratings of each other as mentors and mentees in the context of an online peer-review system supporting accounting professionals to improve their audit and analysis skills.

3 Study Method

We used a modified peer review framework we designed in our previous work [1, 2]. It comprises five stages: writing, feedback by reviewers, rewriting/revising, publishing and back-evaluation of reviews by the authors. In our case, authors are the mentees while reviewers are the mentors. However in this study, our participants are both authors and reviewers. Peers are assigned into review groups of at most four peers based on their competence, measured using an initial calibration task. Each subsequent grouping is based on their competence, as adjudged by their peers, from the preceding peer review session. Because the participants for this study were employed in a full-time work term, they had limited time for completing an initial calibration task. In its place, we used their past course average to measure their competence for the first peer review session. To classify them as strong or weak peers, we calculated the average competence score (i.e., average of their course averages) for all the participants, and every participant with a competence score (course average) less than this value is classified as weak peer, while every participant with a competence score (course average) greater than or equal to the group average is classified as a strong peer.

To frame our analyses, we proposed two null hypotheses:
1. $H_{a1}$: Mentees with high trust scores will be generous in their ratings and offer high ratings in the back-evaluation of their peer mentors.

2. $H_{a2}$: Mentors with high trust scores will be generous in their ratings and offer high ratings in their review feedback.

4 Experimental results and discussion

We conducted a study with professional accountants, who are also registered students of the Master’s of Professional Accounting Program at the Edwards School of Business at the University of Saskatchewan. Thirty-seven (37) students consented to take part in the study. However, only seventeen of them completed the study. The goal was to support them in group peer mentorship to improve their audit, analysis and peer review skills. The study comprised an initial survey to measure the trust score of each participant, using Rotter’s interpersonal trust / distrust scale, and two peer review sessions using an existing online peer assessment system, called Peerceptiv [21], during which the participants analyzed two audit cases and also reviewed each other’s analyses. For both peer review sessions, the peer mentors provided feedback based on three criteria: 1) consideration of engagement issues and risks ($Qn1$), 2) evaluation of accounting choices ($Qn2$), and 3) consideration of materiality, audit approach and audit procedures ($Qn3$).

Participants were assigned into groups of at most four peers for each session. The first assignment was based on the students’ course average, instead of an initial calibration task, because due to their working full time, they did not have time to complete three tasks. In the second peer review session, the participants were re-grouped using their peer ratings from the first peer review session. At the end of each of the two peer review sessions, participants were asked to evaluate their experiences both as authors (mentees) and reviewers (mentors).

As mentioned earlier, we asked the participants to complete an initial short survey by Julian Rotter [12, 25] to measure their interpersonal trust score (see Fig. 1). We found that the average trust scores of the 17 participants that completed the study is 75.76, which Rotter defined as a mild level of interpersonal score, and the standard deviation is 6.52. This trust score is consistent with other research involving financial statement auditors, where the participants were also found to have, on the average, a mild level of interpersonal trust score [22].
To answer our research question, we collected both the individual and average peer review ratings that participants as mentors gave their peers and the back-evaluation ratings that the participants as mentees gave their mentors, all in a series. These values were each compared with their trust scores.

1. **H_{01}**: Mentees with high trust scores will be generous in their ratings and offer high ratings in the back-evaluation of their peer mentors.

   We calculated the correlation between mentees’ (authors’) trust scores and the back-evaluation ratings that they gave their peer mentors (reviewers) (see Table 1). Our results showed that there is a moderate negative correlation (-0.2047) between their trust scores and the back-evaluation ratings that they gave their peer mentors. That is, the higher their tendency to trust, the lower the ratings they give their peer mentors. Although, this result looks contrary to the definition of trust by Mayer, Davis and Schoorman (1995) that a trustor would be willing to be vulnerable and expect to take risk due to the uncertainty and inability to control the other party; it further reinforces the findings of [26, 35] that trustworthiness is not synonymous to gullibility. That is, high trustors tend to have high expectation of other people. In our case, the more trusting mentees have high expectations of the performances from their peer mentors, therefore, they penalized their mentors more when their expectations of their feedback was not met. It further explains why, despite the fact that our participants have on the average a mild level of interpersonal trust (75.06), the average rating that they gave their peers for their reviews was 1.6 out of the total possible rating of 7. We also split the sample into quartiles based on participants’ trust scores. Table 2 shows the mean and standard deviation of the review and back-evaluation ratings for each quartile of the sample.
Table 1, Correlation: trust scores vs. back-evaluation & trust scores vs. review ratings

<table>
<thead>
<tr>
<th></th>
<th>Trust score vs. Back-evaluation</th>
<th>Trust score vs. average reviews ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>-0.2047</td>
<td>0.0208</td>
</tr>
</tbody>
</table>

Table 2, Mean and Standard deviation of trust scores, peer review and back-evaluation ratings for each quartile of the trust scores

<table>
<thead>
<tr>
<th>Quartiles</th>
<th>Trust Scores</th>
<th>Review Ratings</th>
<th>Back-evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean St. Dev</td>
<td>Mean St. Dev</td>
<td>Mean St. Dev</td>
</tr>
<tr>
<td>Q1(72)</td>
<td>70.54  1.82</td>
<td>4.00  1.15</td>
<td>2.08  1.69</td>
</tr>
<tr>
<td>Q2(76)</td>
<td>73.00  3.35</td>
<td>5.15  1.67</td>
<td>2.23  1.79</td>
</tr>
<tr>
<td>Q3(78)</td>
<td>76.92  0.73</td>
<td>5.23  1.12</td>
<td>1.08  1.54</td>
</tr>
<tr>
<td>Q4(92)</td>
<td>84.15  6.02</td>
<td>4.92  1.44</td>
<td>1.31  1.94</td>
</tr>
</tbody>
</table>

These results do not support the hypothesis that mentees with high interpersonal trust scores will be generous and give high ratings in the back-evaluation of their peer mentors.

2. **H2**: Mentors with high trust scores will be generous in their ratings and offer high ratings in the review feedback that they give their mentees.

We calculated the correlation between the average reviews feedback and the reviewers’ trust scores. Our result showed that there is a very weak positive correlation between their interpersonal trust scores and review ratings (see Table 1). That is, a highly trusting reviewer (mentor) will most likely offer high ratings to their author (mentee). We see this as a benevolent act, which [15] described as one of the factors that contribute to the development of trust in a trustee. However, the correlation is weak, which is an indication that more trusting people tend to give slightly positive ratings as mentors to the work of their mentees. We also calculated the correlation between their trust scores and the individual review ratings for each criterion, Qn1, Qn2 and Qn3 that was used in the peer review (Table 3). The results show that Qn1 (0.0344) and Qn2 (0.0186) have a weak positive correlation with their trust scores. Qn3 and trust scores show a very weak negative correlation (-0.0023). We also computed the R-squared values, which show that for each criterion used in the peer review, the trust score is a weak predictor of the ratings that they give their peers as mentors (see Table 3).

Table 3, Correlation: trust scores vs. individual review ratings

<table>
<thead>
<tr>
<th></th>
<th>Trust score vs. Qn1</th>
<th>Trust score vs. Qn2</th>
<th>Trust score vs. Qn3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.0344</td>
<td>0.0186</td>
<td>-0.0023</td>
</tr>
<tr>
<td>R-Squared(R²)</td>
<td>0.0012</td>
<td>0.0003</td>
<td>5.194x10⁻⁵</td>
</tr>
</tbody>
</table>
These findings, therefore, do not support the hypothesis that mentors with high trust scores will always be generous in their ratings and therefore, offer high ratings to their mentees.

5 Conclusion and Recommendation

Group peer mentorship relies on interpersonal relationships that can go wrong and deviate from its objective if not properly managed. Research had identified communication, differing expectations and circumstances, appreciation of how different circumstances affect each person and trust, as some of the factors that can contribute to the success of mentorship [10]. In this study, we investigated the relationship between peers’ interpersonal trust, computed from Julian Rotter’s trust / distrust scale [12], and their rating behavior both as mentors and mentees. The 17 participants that completed this study have an intermediate trust score average (75.06). Our results showed that their interpersonal trust score, which is an indicator of their tendency to trust their peers, is a very weak predictor of their rating behavior as peer mentors, as shown in their R-Squared values. This is a good result, since it shows that even trusting people show no significant positive bias in their rating behavior. However, when the participants were in the role of mentees, they gave to their mentors (reviewers) lower ratings the more trusting they were, i.e., more trusting mentees had a moderate negative bias in the back-evaluations of their peers. These findings show that peers’ rating behavior is influenced by their trust score and it depends on the roles that they assume in the mentoring relationship. The results from this study are preliminary. With further experiment in a large scale, these findings can be useful in explaining the differences in the feedback that peers give themselves in mentorship and collaborative learning.

There are a few limitations to this study. First, we had relatively few participants who gave consent and only a percentage of the consented participants completed the experiment. In the future we will try to run in a larger-scale experiment. Also, due to the small size of the study, it was impossible to run a controlled experiment, which could have helped in testing how other factors, such as group composition and duration of the study, could have affected the results.

6 References

Group Learning, Student Clustering and Peer Mentorship

Donu Anastasiu Popescu$^1$, Ovidiu Domsa$^2$ and Nicolae Bold$^3$

$^1$University of Pitesti, Faculty of Mathematics and Computer Science, Romania,
dopopan@gmail.com

$^2$University “1 Decembrie 1918” University of Alba Iulia, Romania,

$^3$University of Agronomic Sciences and Veterinary Medicine Bucharest, Faculty of Management, Economic Engineering in Agriculture and Rural Development, Slatina Branch, Romania, bold_nicolae@yahoo.com

Abstract: Adaptive learning is an important factor in the normal development of the education of the individual students. Taking into account the differences between the students in a classroom (or even employees in a company), things get complicated in the sense of the distribution of the information for the learners. Solutions to this issue are given to newer or upgraded methods of learning and assessment which are implemented nowadays and their results are quite encouraging. These methods form a system that can be very useful, a system which would comprise techniques such as student clustering, peer mentorship and group learning. In this matter, this paper presents a modality to gather these methods into a system that is applied in some courses in lower grades of high-school or faculty, with results regarding the quality of learning of the students.

Keywords: learning, cluster, group, mentoring.

1 Introduction

The learning process has multiple aspects and various situations appear within this process. One of the most problematic aspect of these is that, in the process of learning in clusters (or in groups), the levels of knowledge of different students may be distinct, which requires adapted methods of learning. Also, perspectives of teacher and students on the educational process are different. Comparative studies on this item can be found in papers [6] and [7].

Thus, frequent situations of this kind appear. For example, at the beginning of a course, the student may not have the minimal knowledge which will be used in the next period. In this matter, the teacher must make knowledge updates or introduce new concepts. These situations are frequent in cases of passing from a tuition cycle to a superior cycle or in cases of perfecting courses, where learners have different ages or come from different institutions.

For example, in the first year of high school, a class may be formed of 5 students from a school, 10 from another and so on. In the majority of the cases, the students that come from the same school have similar levels of knowledge, but it is a major possibility that students from different schools to be trained in a distinct way. The teacher may be constrained to reiterate some concepts. Due to this fact, some students can find the situation dull and unnecessary, being more interested in the applicative part of these concepts, while those who are unaware of these notions have to be trained differently. A useful solution to this problematic situation consists in the
modality of work differentiated on clusters of students. This clusterization is made for the easiness of the educational process and does not imply any other aspect of any kind. Clusters are useful in situations as the one presented, as shown in papers [12] and [13].

To know exactly the situation at the start of the learning period, we use a system which provides as accurate as possible data related to levels of knowledge of students, in order to make the grouping on clusters of students. After this step, the customized data is taught to different clusters of students. This paper presents a model of creating clusters of students that can be later used in the process of learning, using different methods and techniques of education.

We can also say that a characteristic of this model is the adaptiveness, due to the fact that it follows the specific needs of the learner by reporting to the status of known notion (marked by 1) or unknown notion (marked by 0). More specific aspects of an adaptive system in education can be seen in [1].

Also, these kinds of methods of education can be considered as exploratory methods of education, whose characteristics are presented in paper [8]. We called them exploratory because their introduction in education is relatively novel (at least in Romania and surrounding regions) and they have to be studied harder in order for their full efficiency to be shown.

2 Student clustering: model description and algorithm

Based on some variables of the student, a general level of knowledge of a student can be established based on the school of origin and the results of an initial test made by the teacher.

In this matter, we will present a model which solves the problem presented in the introduction. In this model, we will use some notations. These are:

- \( N_k \): the number of the students in the classroom;
- \( K \): the number of institutions of origin of the students in the classroom;
- \( S_{C_1}, S_{C_2}, \ldots, S_{C_K} \): the institutions of origin of the students from the classroom;
- \( n_{s_1}, n_{s_2}, \ldots, n_{s_K} \): number of students that come from the institution \( S_{C_1}, S_{C_2}, \ldots, S_{C_K} \);
- \( T_1, T_2, \ldots, T_K \): the set of unknown concepts of the students from the school \( S_{C_i}, S_{C_2}, \ldots, S_{C_K} \) found out after a set of questions answered at the start of the course; these variables can consist in two values, for a given notion \( W_i \):

\[
T_i = \begin{cases} 1, & \text{if } W \text{ is known by the group } s_{i}, \text{ or } i = 1; K, \\ 0, & \text{otherwise} \end{cases}
\]

- \( W_1, W_2, \ldots, W_K \): the set of concepts that are contained in the curricula, where \( P \) is their number.

The sets \( n \) and \( T \) are found in the relations presented in (1) and (2).

\[
\begin{align*}
N_{SE} &= n_{s_1} + n_{s_2} + \cdots + n_{s_K} \\
T &= T_1 \cup T_2 \cup \cdots \cup T_K = \{0, 1\}
\end{align*}
\]

This level of knowledge (LK) for each student is measured as an average between the results on the test (also measured as a ratio and denoted by TR) and the degree of confidence (DC) of the institution of origin, as shown in (3).
\[
L_i = \frac{0.75 \times TR_i + 0.25 \times DC_j}{0.75 + 0.25} , i = 1 \text{, } N_{st} , j = 1 \text{, } K
\]

\[
TR_i = \frac{\text{no of known concepts } (W_z \text{ where } T_y = 1)}{p}, z = 1 \text{, } P, y = 1 \text{, } K
\]

TR is calculated as a ratio between the number of known concepts at the test and the total number of concepts. The number of known concepts is considered the number of concepts for which \( T \) for the student \( i \) which is part of the group \( ns_y \) is equal to 1.

The degree of confidence of the institution can be related in this way:
- for educational institutions (e.g. schools, faculties), the degree of confidence can be considered as a ratio between the number of graduates and the number of enrolled students;
- for companies, any indicator that shows the efficiency of the personnel (e.g., the proportion of submitted projects related to the total number of projects) can be taken into account.

In order to work in an easier and more attractive manner with the students, these will be grouped in two clusters M and P. The cluster M is formed of the students that must receive more detailed both practical and theoretical information related to certain notions established in the set T. The other cluster, P, will contain the students that have learned the theoretical aspects of the concept \( T_i \) from the set T and want to deepen and apply into practice this concept.

The reason for choosing two clusters instead of a larger number of clusters is given by the easier focus of the assessor on two clusters. Besides, related to a piece of information, the learner has two possible states: (a) know or (b) not know the information. Another reason is that the process of peer-mentoring would be clearer if two clusters are involved.

The established variables used in this model will have to be used in different methods in order to obtain the desired results.

Firstly, the students are tested, for establishing the variables \( N_{sk} \), \( K \), \( SC_1 \), \( SC_2 \), \( SC_3 \), ..., \( SC_k \), \( n_1 \), \( n_2 \), ..., \( n_k \), \( T_1 \), \( T_2 \), ..., \( T_k \). The test is built in order to ease the creation of the clusters. In this matter, the models of testing described in papers [2–4] or in paper [5] can be used. Based on certain keywords given at the beginning, the teacher can select some generated questions after certain restrictions are set. Then, \( L_i \) for each student is calculated.

After this step, the clusters are formed based on \( L_i \). After certain set periods of time, the teacher can assess in a customized way the students using the same system of testing based on keywords. The final result would consist in reaching a balance between the two clusters, reducing the information gaps between those two clusters, assuring the performance of the cluster P and reaching a good level for the cluster M.

The clusters are formed dynamically. This means that:
- for every piece of information or notion W, the clusters can change;
- the whole group may be considered a specific type of group (e.g., a team that participate at a contest), then the clusters are formed analogous.

The graphical presentation of the system can be seen in Figure 1.
In order to ease the work of the teacher, an algorithmic solution exists to solve this model. Practically, the model is transported in a mathematical language. The steps followed by this algorithm are presented next.

**Step 1.** Input data is read. Input data consist in the results to the test and their interpretation.

**Step 2.** The students are grouped in the format of $n_s_1$, $n_s_2$, ..., $n_s_k$ based on the results on the initial test.

**Step 3.** The students from each group are verified if they know the concept $w$ from the set $T$. If any student wishes to recapitulate the concept $w$, it is introduced in the group $M$.

**Step 4.** The cluster $P$ is determined as a difference between the set of students and the cluster $M$.

**Step 5.** The two clusters are formed.

Practically, it can be summarized in the next pseudocode:

```plaintext
for each $w \in T$ do
    $M \leftarrow \Phi$
    for $i = 1, K$ do
        $M \leftarrow M \cup S_i$
        for each $j \in S_i$ do
            //if student $i$ from $S_i$ wishes to recapitulate the notion $w$, although
            //the results showed that the student knows the notion $w$
```
if \( j \in S \) wishes to recapitulate the notion \( w \) then

\[
M \leftarrow M \cup \{j\}
\]

endif

endfor

endfor

\( P \leftarrow N_S \setminus M \)

write \( w, M, P \)

endfor

Practically, the algorithm can be schemed in the form shown in Figure 2.

Fig. 2: The scheme of the algorithm

3 Group learning

Learning can be a dull activity if methods used are not interesting. That is, a method to improve learning efficiency is group learning. This method can have two forms:
- classroom group learning (CGL), made in the classroom; the students from the same cluster can relate to each other in the classroom;
- online-based group learning (OGL), which would continue indifferent to the time and which will be presented in the next lines.

OGL uses structures and methods such as databases, keywords or searching operations in order to update the actual information and offer valid information to persons which access the system.

The structure used in OGL is the piece of information (PI), which is formed of two components: a keyword \( T \) and the actual information \( I \). Practically, PI is a pair of a keyword \( k_W \) and a piece of information \( I \) (PI = \([k_W, I]\)). We will also use the variables:
- \( N \): the number of pieces of information PI, \( i = 1, N \)
- SPI: the PI searched by the user.

The model of functioning is a simple one and it will be presented in the next rows. We must mention that an initial database of PI is established in time. An illustrative scheme of the system can be seen in Figure 1.
Fig. 1: The OGL method proposed by the authors

Firstly, a keyword desired to be searched in the system is set. Information about this word (whether it can be found in the database, how many PIs contain the keywords etc.) is output. Every PI that contains the keyword is output.

If at the previous phase the information was not enough, complete or the user does not find them useful, the new piece of information is search externally (on the Internet, on books, papers etc.), then these pieces of information are introduced in the system in the format provided by the system: [keyword, piece of information].

If the user does not find any information related to the searched keyword, then pieces of information are searched externally, then systematized, completed with examples and introduced in the system in order to be used by other members of the group.

Another method of introduction of pieces of information is the introduction of an admin or a mentor who has the responsibility to avoid the intrusion of incorrect or redundant information and to fill will competent information the system whether it is needed. An important facility of the system is represented by a component which notifies the users that a new piece of information was added in the system, whether is the case.

The modality of functioning of the system is presented in Figure 2.
4 Peer mentorship: learning by teaching

As said at the end of section 3, OGL is kind of an informal peer mentorship. It is informal because peers interact only through online means and the formal rules of mentorship are not fully kept. Thus, a more formal type of mentorship should be used, in which students from cluster P are mentors and those from cluster M are mentees. This is made under the strict supervision of the teacher.

Peer-mentorship has been declared to have beneficial effects on learning. According to a study presented in paper [9], the peer mentorship led to the increase of submitted projects for junior level researchers, but also to the enhancement of the focus on more specific projects for the chosen specialty in case of each researcher.

Why is peer mentorship beneficial? Researches show that five instances of a peer make this method successful: connecting link, peer leader, learning coach, student advocate and trusted friend [10]. With all the risks that come from this method (vulnerability, non-acceptance of mentors etc.), this is known to be beneficial for the main reason that feedback from students received by the mentor is more honest and clearer than the feedback received directly by the teacher, according to the same source.

The peer mentoring in case of our system is simple: the students from the cluster P are trained to mentor those from cluster M. In this way, participants from both clusters would profit: those from cluster M teach in a different way the needed notions, while those from cluster M form a more solid knowledge base for the future notions, thus, they learn by teaching to others.
5 Unifying the three methods: system description

The three methods presented in the previous section can be united in a system that would consist in a powerful tool used in the classroom. Some methods of the system are already used in programming classes within the National College “Radu Greceanu” from Slatina (e.g., formation of clusters, peer mentors). Basically, an extended form of the system would behave in this way:
- the clusterization of the students is made at the beginning of a period;
- after the clusters are formed, the student can learn in groups or clusters (M and P), using both CGL and OGL;
- parallel with the group learning, the peer mentorship can be an important tool in the system.

A scheme of the system that would unite the three methods presented above is presented in Figure 3.

![Fig. 3: Unifying the methods in the system](image)

The methods can be used in complementarity or in different combinations and can led to good results, especially regarding the qualitative aspect of learning. This means that, within a classroom, the difference between the clusters is decreasing, creating a balance between the two clusters from the skills gain point of view. As student in high-school, one of the authors had been benefited by the benefits of a type of this system.

6 Results

The model is used in classrooms with mathematics-informatics profile in high schools, at an empirical level. We shall take an example of a high school classroom of
9th grade at the Web Programming class. The class is made from 25 students. The number of institutions will be 7 (K=7), thus the number of groups (ns) will be 7. LKs for each student based on three notions W1 (the usage of Internet), W2 (the tag elements in HTML) and W3 (creation of simple websites) are shown in Table 1, considering that all the students are aware of notion W5. The DC of every institution is calculated by dividing the average of the means of all students from the respective institution to 100. The means are obtained at the exam of National Evaluation, which is given in the 8th grade [11].

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>LK</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.47</td>
<td>0.72</td>
<td>0.72</td>
<td>0.47</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>ALK</td>
</tr>
<tr>
<td>LK</td>
<td>0.43</td>
<td>0.71</td>
<td>0.71</td>
<td>0.41</td>
<td>0.43</td>
<td>0.43</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.43</td>
<td>0.43</td>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Levels of knowledge for each student

The average LK (ALK) is calculated as an average of LKs of all 25 students. In our case, the average has the value of 0.55. This value divides the students in the two clusters in this way: if LK_i < ALK, then the student i is in the cluster M and if LK_i > ALK, the student i is in the cluster P. The groups depending on institutions are found in Table 2. The concept W2 will be considered the tag elements in HTML. Some of them learned in the general school this notion and reached the next one W3 (creation of simple websites).

<table>
<thead>
<tr>
<th>Institution (SC_i)</th>
<th>DC</th>
<th>Number of students (ns)</th>
<th>T1 for W2</th>
<th>Cluster P</th>
<th>Cluster M</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Eugen Ionescu&quot; School</td>
<td>0.90</td>
<td>12</td>
<td>YES / 1</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>School no. 7</td>
<td>0.74</td>
<td>2</td>
<td>NO / 0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>&quot;Radu Greceanu&quot; National College</td>
<td>0.84</td>
<td>2</td>
<td>YES / 1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Ianca School</td>
<td>0.66</td>
<td>1</td>
<td>NO / 0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>School no. 3</td>
<td>0.72</td>
<td>2</td>
<td>NO / 0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>&quot;Constantin Brâncoveanu&quot; School</td>
<td>0.76</td>
<td>4</td>
<td>NO / 0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Sport-Scheduled High school</td>
<td>0.74</td>
<td>2</td>
<td>NO / 0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>0.77</td>
<td>25</td>
<td>-</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 2: Clustering a classroom depending the notion W2

The cluster P will be formed of 12 students, after the initial test is given and the results are interpreted. The two students coming from "Eugen Ionescu" school wanted to reiterate the notion T1, thus will be considered in cluster M. The students from the cluster P will learn to develop simple websites, while the students from cluster M will
learn the tags from HTML, in a future the two clusters being assessed accordingly to the knowledge gained, both clusters being brought to a good level of knowledge in a certain amount of time.

7 Conclusions

The LMSC shows an adaptive model of teaching, according to the needs of the student. Thus, the separation made in the two clusters has an educational purpose, the learning having a continuous flow for every individual student. We must also mention that for every notion $T_p$, clusters M and P change, thus these clusters are dynamic. This method is also based on the native talent of the teacher and of his managerial and organizational competences, his talent compensating the weaknesses of this method (determined by the projection of the model or the methods used for testing). A future work would be the integration of this model into a more complex model of assessment and teaching which is currently built.

8 References


Usability of an Active Video Watching System for Soft Skills Training

Lydia Lau¹, Tanja Mitrovic², Amali Weerasinghe³, Vania Dimitrova¹

¹University of Leeds, United Kingdom
²University of Canterbury, New Zealand
³The University of Adelaide, Australia

Abstract. This paper reports an empirical study on the usability, including acceptance, of the Active Video Watching system (AVW) for learning advanced presentation skills. The AVW provides a Personal Space for individual learners to link their personal experiences while watching the selected videos. The comments collected can be used by the individuals to reflect on their own thoughts or to be shared with other learners in the Social Space. This sharing environment enables the learners to see if their own views concur or differ from the others. The tutor can also obtain an overview quickly to decide on an appropriate learning intervention if needed. The paper presents the initial results, focusing on the usability of AVW Personal Space and Social Space. Implications for improvement and plans for further analysis of the data are outlined.

1 Introduction

Soft skills (such as communicating, negotiating, problem solving, collaborating) are key 21st Century skills, and are crucial to improve employability in the knowledge economy. This research investigates the use of augmented social interactions with videos to facilitate soft skills development by promoting reflective learning linked to real world experience. Within the EU project ImREAL[1], an Active Video Watching (AVW) system was developed. The prime objective is to support reflective experiential learning by providing personal and social spaces for interactive watching and annotating of videos. AVW was trialled in real contexts with students and tutors in several studies. This showed positive reactions from the tutors as a potential addition to their teaching tools for soft skills, such as communication, negotiation, leadership. Specifically, tutors appreciated the opportunity to reuse existing videos from YouTube as part of their soft skills training practice. Using videos for teaching is particularly appealing for the millennial learners who are technology savvy and use a vast amount of interactive media in social contexts, as a means of conversation and expression. Video-based learning is seen as one of the main strategies to provide engaging learning environments for millennials (together with gamification and mobile learning) [2].
Pedagogically, it remains a challenge for tutors to find an effective way to embed these learning resources within the context of ill-defined domains [3], and enable learners to be critical and reflective. This paper uses part of the results from a follow on study with postgraduate students from several universities using AVW as an informal learning tool for advancing their presentation skills. The study examines how learning can be scaffolded [4] by using a video sharing space such as the AVW and the extent of supporting the four modes of overt engagement: Interactive > Constructive > Active > Passive, as espoused from the ICAP framework [5].

The chosen soft skills domain is delivering a pitch presentation. Increasingly, researchers are asked to present their work in a short, sharp and engaging manner. This is also crucial in businesses where a new product or a proposal has to be presented to customers or funders. Pitch presentations are being used as a form of public engagement vehicle, which aim to excite, persuade, and opening up opportunities. This can be at odds with the usual presentations for research dissemination which would examine the rigour of approach, grounded in the past for credibility, and the technicalities of the techniques employed. The overarching purpose of the learning material used in AVW is to help postgraduates acquire new skills for a pitch presentation, and use this opportunity to reflect on their own presentation skills.

There are assumptions in this approach:

- past experience may be recalled while watching a video that is useful for personal reflection – i.e. when using the AVW Personal Space;
- sharing these experiences may be useful for opening up the learner’s mind when there is dissimilarity amongst the experiences, or improving the learner’s confidence when similar experiences were voiced by others – i.e. when using the AVW Social Space.

This paper examines the validity of these assumptions with the use of the AVW by surveying the learners on the acceptability and usefulness of the features provided by the personal and social spaces in the AVW. The work is part of an ongoing collaboration between the University of Leeds (United Kingdom), the University of Canterbury (New Zealand), and the University of Adelaide (Australia), which aims at scaling up the deployment of AVW for soft skills training, and its potential in the development of interactive personalised nudges for self-regulated learning.

2 The Active Video Watching System (AVW)

The AVW is a controlled video watching environment for the facilitation and collection of video comments from the learners, and ultimately more beneficial for reflective activities and assessments. The technical platform for AVW development was .NET, SQL Server, AJAX, Visual Studio 2010 and Telerik web controls; and using YouTube APIs. AVW enables a tutor to:

- Create an interaction space for active video watching aligned with the purposes of their training;
• Upload selected videos from YouTube and add short video descriptions;
• Define main aspects that can direct the learners’ attention to specific points related to the videos. Each aspect is presented with a term, or short phrase, for scaffolding learner’s experience. Aspects guide a learner to associate a comment with a particular learning concept in the video. The system automatically captures not only the comment but also the place in the video (i.e. the time elapsed from the start) when the comment is made; and
• Approve the learner comments for sharing with others.

The tutor can also download the interaction data from the AVW as an XML file that includes the comments tagged with the aspects and the timing in the video when a comment was made. The file can be processed for further analysis to get deeper insights into the learners’ experience with AVW. In the previous trials of AVW, the tutors did not use this data. Researchers within the ImREAL project processed some of the data to identify the focus of attention for individuals or group of learners [6].

The main functionality of the AVW from a learner’s perspective is explained in the following sections.

2.1 Personal Space

A learner can access one or more videos within his/her ‘Personal Space’. While watching a video, the learner can pause at any time to record a comment (see Fig. 1). Each comment will be time-stamped when the video is paused for the text entry with any associated aspect (from the list of aspects defined by the tutor). For example, the pitch presentation study presented below uses the following aspects to put self-regulation context around the comments the learner makes on tutorials for soft skills:
• “I am rather good at this”;
• “I did/saw this in the past”,
• “I didn’t realise I wasn’t doing this”; and
• “I like this point”.

The learner can exit and re-enter a video anytime, and previously made comments can be viewed at the bottom of the page.

2.2 Social Space

Comments made by learners will only appear in the Social Space after the tutor has approved their release. Learners can enter the social space and glance through the comments made by others (anonymised) together with own comments for each video (see Fig. 2).
The learner can sort the comments according to:

- elapsed time – for identification of interesting places in the video that have attracted a number of comments, and to check similarity or otherwise with own comments (e.g. see around 33 secs in Fig. 2, the learner felt it was interesting, concuring with another learner who said “very nice visual”);
- aspect – to see the extent and kind of experiences shared around the same learning concepts.

The learner can also ‘rate’ each comment, according to the prompts setup by the tutor, for further reflection (see popup box in Fig. 2).

In addition to reading/rating the comments, the learner can click on ‘view video snippet’ and watch the part of the video that the comment refers to.
3 Experimental Study with AVW

This study was conducted to understand how learners would use both the Personal Space and Social Space for informal learning of how to deliver a ‘pitch’ presentation.

Participants: Email invitations were sent to the mailing lists of PhD communities in English speaking countries that the authors had access to, with the expectation that general presentation skills were already acquired.

Set up: Learning objects were a set of eight videos, carefully selected from YouTube that covered different aspects of presentation skills – four tutorials and four examples. Criteria for their selection were: (i) appropriate content (covering the spread of opening and closing, structure, delivery, and visual aids; or examples of pitch presentations); (ii) reasonably short (no longer than 10 minutes); (iii) with a balance of gender for the presenters; and (iv) two good examples and two not as good.

Wording for aspects for reflection are designed to encourage the learners to put their comments within selected learning context. For the tutorials, the aspects provided were: “I am rather good at this”, “I didn’t see this in the past”, “I didn’t realize I wasn’t doing it”, and “I like this point” to stimulate learners to recall / relate to their own experiences. For the example videos, the aspects provided were: “Delivery”, “Speech”, “Structure”, and “Visual aids” – concepts that were covered in the tutorials.

In the social space, the ratings provided were: “This is useful for me”, “I hadn’t thought of this”, “I didn’t notice this”, “I don’t agree with this”, and “I like this point” – to promote deeper level of reflection.

Data collection methods: Three survey questionnaires were designed to collect user data, and to set learning tasks for pre- and post-test analysis.
Survey I: participant’s profile such as demographic information, background experiences, motivation and attitudes; then a series of questions relating to participant’s knowledge of presentations; and his/her action plan for preparing and presenting a pitch presentation.

Survey II: same questions for participant’s knowledge of presentations and an update of action plan for preparing and presenting a pitch presentation; NASA-TLX instrument [7] to check participant’s perception of cognitive demand when using AVW Personal Space; Technology Acceptance Model (TAM) [8] to check the participant’s perceived usefulness of Personal Space for informal learning of presentation skills; and questions on usability of the AVW Personal Space.

Survey III: same questions for participant’s knowledge of presentations and an update of action plan for preparing and presenting a pitch presentation; NASA-TLX and TAM for the Social Space; and finally questions on usability of the Social Space.

Procedure for participants:
- complete preliminary learner profile and baseline survey I;
- phase 1: use of personal space to view and comment on the eight videos;
- complete post-personal-space survey II;
- phase 2: use of social space to view other comments and rate if appropriate;
- complete post-social-space survey III.

4 Results and Analysis

50 participants accepted the invitations. 38 completed both phases of the study that spanned across 2 weeks. A prize draw was provided to compensate the time spent by participants who completed the study. Participants were PhD students mainly from New Zealand and the UK, with some from other parts of the world.

AVW usage overview: Table gives an overview of the usage of AVW in both the Personal Space (where the participants watch videos and made comments) and Social Space (where the participants read and rated comments by others).

<table>
<thead>
<tr>
<th>Video</th>
<th>Video Length</th>
<th>Comments [Personal Space]</th>
<th>Comments with ratings [Social Space]</th>
<th>Ratings [Social Space]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutorial 1</td>
<td>2.54'</td>
<td>97</td>
<td>95</td>
<td>603</td>
</tr>
<tr>
<td>Tutorial 2</td>
<td>7.37'</td>
<td>124</td>
<td>123</td>
<td>382</td>
</tr>
<tr>
<td>Tutorial 3</td>
<td>6.55'</td>
<td>121</td>
<td>118</td>
<td>402</td>
</tr>
<tr>
<td>Tutorial 4</td>
<td>6.22'</td>
<td>94</td>
<td>91</td>
<td>261</td>
</tr>
<tr>
<td>Example 1</td>
<td>3.23'</td>
<td>82</td>
<td>82</td>
<td>272</td>
</tr>
<tr>
<td>Example 2</td>
<td>8.28'</td>
<td>100</td>
<td>98</td>
<td>281</td>
</tr>
<tr>
<td>Example 3</td>
<td>6.48'</td>
<td>106</td>
<td>103</td>
<td>283</td>
</tr>
<tr>
<td>Example 4</td>
<td>3.25'</td>
<td>66</td>
<td>62</td>
<td>222</td>
</tr>
</tbody>
</table>
This paper discusses one part of the surveys, namely the degree of technology acceptance and perceived usefulness of the AVW – i.e. on the use of Personal Space, the Social Space, the aspects and ratings for learning advanced presentation skills.

4.1 Technology Acceptance – Quantitative Analysis

The following scale was used for the TAM questions: 1. extremely likely; 2. quite likely; 3. slightly likely; 4. neutral; 5. slightly unlikely; 6. quite unlikely; 7. extremely unlikely. Table 2 summarises the means and standard deviations of the replies.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Personal Space</th>
<th>Social Space</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q1:</strong> I think I would like to use the AVW frequently</td>
<td>3.03 (1.59)</td>
<td>3.68 (1.85)</td>
</tr>
<tr>
<td><strong>Q2:</strong> I would recommend the AVW to my friends</td>
<td>2.76 (1.51)</td>
<td>3.53 (1.98)</td>
</tr>
<tr>
<td><strong>Q3:</strong> Using the AVW would enhance my effectiveness when developing soft skills</td>
<td>2.50 (1.18)</td>
<td>3.08 (1.76)</td>
</tr>
<tr>
<td><strong>Q4:</strong> I would find the AVW useful in my studies/job</td>
<td>2.46 (1.19)</td>
<td>3.26 (1.84)</td>
</tr>
<tr>
<td><strong>Q5:</strong> I would find the AVW easy to do what I want it to do</td>
<td>2.76 (1.34)</td>
<td>3.42 (1.72)</td>
</tr>
<tr>
<td><strong>Q6:</strong> My interaction with the AVW would be clear and understandable</td>
<td>2.73 (1.41)</td>
<td>3.28 (1.58)</td>
</tr>
<tr>
<td><strong>Q7:</strong> I would find the AVW easy to use</td>
<td>2.49 (1.48)</td>
<td>3.13 (1.49)</td>
</tr>
<tr>
<td><strong>Q8:</strong> If I am provided the opportunity, I would continue to use the AVW for informal learning</td>
<td>2.47 (1.33)</td>
<td>3.50 (1.90)</td>
</tr>
<tr>
<td><strong>Q9:</strong> Using the AVW would enable me to improve my soft skills quickly</td>
<td>2.59 (1.33)</td>
<td>3.24 (1.84)</td>
</tr>
<tr>
<td><strong>Q10:</strong> Using the AVW would improve my performance considering the development of soft skills</td>
<td>2.47 (1.11)</td>
<td>3.24 (1.81)</td>
</tr>
</tbody>
</table>

The figures showed that personal space was more acceptable than the social space. However, with the comparatively high standard deviations, the responses were mixed. Qualitative analysis on the comments from the participants gives some explanations.

4.2 Perceived Usefulness – Qualitative Analysis

Questions were asked to target at specific features of the AVW, the general feedback are summarised below.
Personal Space

Usefulness of pausing a video to write a comment
Positive: Can write down the thought straight away before you forget it, link it to a point in time in the video; a good way to produce a personalized summary of a video; It keeps the listener active and alert; It consolidates the thought behind the comment, and forces the user to articulate it, which makes it more memorable.
Negative: Does not allow selecting multiple aspects to a comment.

Usefulness of specifying aspects when composing comments
Positive: Helps identify the category for the comment; it is easy for reviewing comments.
Negative: More than half of the participants did not find this useful; more so for the aspects in the tutorials.

What was seen as most exciting about the Personal Space in AVW
- “someone has already gathered fairly good tutorials, saving me the time to search youtube”
- “Ability to move through lessons at own pace”
- “Simplicity and rich and it makes me feel like this is my world immediately I login in. And I can equally makes some comments by editing which is very awesome”
- “I have absolute control, making useful comments intermittently”

What disappointed users about the Personal Space in AVW
Several participants pointed at interface issues, indicating that there was a need for some training to use the system effectively. A number of participants replied “nothing” when asked what disappointed them.

As a whole the users appeared positive and excited about the Personal Space, seeing this as an enabler to write their thoughts on the videos, see personal summaries to make meaning from the video, keeps them active and engaged. However, the self-regulation scaffolds, i.e. the aspects used for video annotation, were not perceived as very useful (especially for the tutorials). Further analysis of the interaction with the videos is required to find out what proportion of the comments did not have aspects associated with them, what aspects could have been missed. It has to be noted though that the experiment followed a rather opportunistic approach, mimicking the interaction with videos in widely used social platforms, e.g. YouTube, and did not provide any specific explanation for the aspects and how to use them. Our goal is to investigate whether such 'organic’ un-structured way of video watching, which can be scaled easily in a variety of informal learning contexts for soft skills, would provide means for user engagement and can trigger self-regulation processes. Our further analysis (see below) is aimed at gaining deeper insights into this.
Social Space

Usefulness of reviewing the comments on the videos made by others
Positive: Just skimming through the comments would be useful; it is seen as a means to check if there is anything missed; Good to get different views/opinions; Reinforce learning: “More useful than I originally thought”.
Negative: Rating were not useful for some participants; Too many comments to go through; Some participants did not find the Social Space useful.

Usefulness of rating the comments of others
Positive: Good as optional, tended to agree with other’s comments; helped to reflect further.
Negative: some participants did not notice the ratings; about a third of participants did not find it useful.

What was seen as most exciting about the Social Space in AVW
- “Share of different knowledge”
- “Some of the comments were very funny”
- “Ability to rate/categorise comments”
- “The videos and the reflections were very helpful”
- “Easy to use”

What disappointed users about the Personal Space in AVW
- “The user interface - page refreshing to the top”
- “Asynchronised communication with other viewers, rating without discussion, actually, there is no communication at all”
- “Too much repetition, too many useless comments”
- “Comments are not grouped by similarity”
- “Very crude tool to try to manage complex learning processes”

The participants’ opinions about the usefulness of the AVW Social Space were less positive than for the Personal Space. This can partially be explained with the lack of guidance to explain what the pedagogic value would be. Despite this, it is reassuring that many participants saw the Social Space as a means to reinforce learning, see alternative points, ensure that something is not missed. There is clearly a need for further improvement which requires computational means to rank/order the comments and to bring the most relevant comments to the users (e.g. based on their profile or interaction behaviour). Despite the unenthusiastic views on the ratings, there is high number of ratings provided (see Table 1). Further analysis of the ratings and comments can shed light how to facilitate the rating process to realise its value for self-regulated learning (e.g. the learners can be nudged to read/rate comments).
5 Conclusion and Future Work

The study sheds light on the validity of our assumptions and points at the need of further analysis.

The ability for learners to pause a video and make comments in the AVW Personal Space helped notes to be made about their experience with the videos at different points. These comments could be retrieved easily for later use. The passive form of scaffolding, in the form of aspects, did not seem to achieve their aim to trigger recall of past experiences effectively. However, further analysis on individual learner’s comments can be used to get deeper insights into the learners’ engagement and to investigate whether individual differences have an effect on this. This will allow designing appropriate strategies for personalised intervention.

The use of the AVW Social Space for shared experiences was valued only by some participants although the amount of rating was quite high. This may be due to the lack of meaningful learning task being set and/or lack of training on how the features could be used (as the usability requires improvement). The overwhelming volume of comments call for some form of intelligent filtering for individual learners. One possible way forward is to analyse learner’s comments on the videos to identify appropriate comments to be selected in the Social Space. This could be based on learner’s sentiments, topics covered, or number of reflection words.

Overall, the intelligent scaffolding can provide mentor-like features to trigger self-regulated learning for developing of soft skills from active video watching contexts.

References

Workshop 4

2nd International Workshop on Affect, Meta-Affect, Data and Learning

(AMADL 2016)

https://sites.google.com/site/iwamadl2016/
2nd International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2016)

Benedict du Boulay - University of Sussex, UK
Beate Grawemeyer - Birkbeck University of London, UK
Manolis Mavrikis – University College London, UK
Genaro Rebolledo-Mendez - University of Veracruz, Mexico
Olga C. Santos - aDeNu Research Group - UNED, Spain

There have been many attempts to take emotions and affect into account during the design and the deployment of ITSs. The evidence for the consequential impact on learning is beginning to strengthen but the field has mostly focused on the complexities of affect and emotion detection, recognition and consequential system reaction. Going beyond this, there are indications that meta-affect —knowledge about self-affect —also plays a role. This workshop (AMADL 2016) focuses on how tutoring systems should detect and respond to the learner’s meta-cognitive and meta-affective skills and knowledge. The four papers accepted cover different aspects of these issues.

We thank the Programme Committee for their work.

• Vincent Aleven, Carnegie Mellon University, Pittsburgh, USA
• Ivon Arroyo, Worcester Polytechnic Institute, USA
• Roger Azevedo, North Carolina State University, USA
• Amy Baylor, Amy Baylor Consulting, USA
• Jesus G. Boticario, aDeNu Research Group. UNED, Spain
• Rafael Rojano Cáceres, University of Veracruz, Mexico
• Art Graesser, University of Memphis, USA
• James Lester, North Carolina State University, USA
• Rose Luckin, University College, London, UK
• Erika Martinez-Miron, Universidad Politécnica de Puebla, Mexico
• Ma. Mercedes Rodrigo, Ateneo de Manila University, Philippines
On the Feasibility of Providing Real-Time Adaptive Support for Motivation and Emotion in Intelligent Tutoring Systems

Seth A. Martin, Joseph F. Grafsgaard, Nicholas V. Mudrick, Michelle Taub, Roger Azevedo
North Carolina State University, Department of Psychology, Raleigh, NC, USA
(samarti7, jfgrafsg, nvmudric, mtaub, razeved)@ncsu.edu

Abstract Few intelligent tutoring systems (ITSs) currently utilize multichannel data to foster cognitive, affective, metacognitive, and motivational (CAMM) self-regulatory processes, as well as adaptively adjust to learners’ capacities to flexibly regulate their emotions. However, accurate modeling of motivation and emotion during learning poses a great challenge, mainly due to how the construct is conceptualized, measured, and analyzed leading to severe constraints for researchers interested in treating motivation as a dynamic process that fluctuates during learning. In this study, 194 college students used MetaTutor, an ITS for human biology. The experimental design included a battery of self-report measures and trace data to assess learners’ CAMM processes during learning with MetaTutor. We specifically investigated learner achievement goal orientation vis-à-vis the Achievement Goal Questionnaire-Revised (AGQ-R). Seven clusters of goal orientation were identified using Expectation-Maximization. Learners in two clusters associated with high mastery-approach goal orientation learned, on average, 52% more than learners in the other five clusters. These learners also exhibited greater emotion regulation capability as shown by their results on the Achievement Emotion Questionnaire (AEQ). These results suggest that learners with distinctly higher self-reports of mastery-approach goal orientation exhibit greater intrinsic motivation and emotion regulation capabilities. The results of this study have important implications for designing adaptive and personalized ITSs that emphasize real-time motivation and emotion regulation during complex learning.

Keywords: Motivation, Emotions, Complex Learning, Intelligent Tutoring Systems (ITS), Self-Regulated Learning (SRL), Adaptive Scaffolding

1 The Importance of Goal Orientation and Academic Emotions for Adaptive Support with an Intelligent Tutoring System

For learners to achieve superior learning outcomes during complex learning with intelligent tutoring systems (ITSs) they must accurately monitor and adaptively regulate their cognitive, affective, metacognitive, and motivational (CAMM) processes [1]. Complex learning requires sophisticated strategies like the synthesis, integration, and coordination of materials [2], the accurate use of cognitive and metacognitive processes (i.e., self-regulated learning [SRL], CAMM: [3] [4]), the ability to flexibly regulate emotions [5], as well as the motivation to persevere through obstacles encountered while learning challenging and complicated materials [6]. SRL has proven adfa, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011
to be an effective strategy to help learners engage in complex learning. For example, results from [7] suggest that learners who engage in SRL strategies are more likely to obtain deeper understanding (e.g., greater learning gains) of complex information than learners who only attempt to acquire declarative knowledge. While motivation has long been known to be an important component of SRL [1] [8]; and one of the core components of CAMM, the examination of motivation with respect to complex learning has been largely ignored by cognitive, affective, educational, learning, and computational scientists. This is perhaps due to the complex nature of motivation and the constructs that define it, as well as the difficulties in detecting, tracking, and scaffolding motivation during learning with an ITS.

More specifically, we argue that the lack of research on motivation by the ITS community is largely due to challenges related to how the construct is conceptualized, measured, and analyzed, which lead to severe constraints for researchers interested in treating motivation as a dynamic process that fluctuates during learning. For example, motivation tends to be operationally defined as a trait and thus stands in stark contrast to conceptualizations of cognition, metacognition, and affect, as states that fluctuate during learning, problem solving, performance, etc. In addition, motivation is typically measured with self-report measures (e.g., MSLQ) along several dimensions (e.g., intrinsic motivation, task value, self-efficacy; based on particular theoretical orientations) where learners are classified according to a particular dimension (e.g., mastery orientation) and assumed to retain and adopt that particular orientation through the entire learning session. Like other contemporary research (e.g., [4] [9]), we challenge this view and adopt the stance that motivation can be re-conceptualized as a state-based process and therefore incorporated as a key component to SRL (in addition to CAM) during learning with ITSs. We therefore make the following assumptions: (1) motivation is a macro-level process that is comprised of several micro-level processes (e.g., goal orientation, task value, self-efficacy) that can be measured with self-report measures to assess learners’ self-perceptions at several key points in time (e.g., prior to, during, and following learning); (2) micro-level motivation processes leave traces that can be inferred from multichannel data during learning (e.g., changes in the frequency and accuracy of use of sophisticated cognitive learning strategies); (3) fluctuations in motivation (e.g., motivational states) are likely to occur throughout the learning session (based on a myriad of contextual factors, internal and external demands, accuracy in monitoring cognitive, affective, and metacognitive processes, perceived value of artificial agents’ scaffolding, etc.); and: (4) if 1-3 are correct then we can design and adaptively scaffold motivational processes based on the use of multichannel data. This study presents a first step towards addressing these assumptions and making ITSs more sensitive to motivational processes during complex learning.

Learner motivation can be understood in terms of the learner’s willingness to engage with a learning environment (i.e., feelings of self-efficacy in using sophisticated strategies, persistence in navigating the non-linear system to search for relevant information, understanding the value of the learning task), to set suitable goals (i.e., setting proximal, efficient, and reasonable goals that lead to reaching the overarching [distal] goal), to devote appropriate cognitive and affective processes (i.e., by appraising the relevance of obstacles and determining if they are critical to their goal and by setting upper bounds of workload expenditure for tasks), and to persist when they encounter difficulties (i.e., how much effort they expend is dependent upon many
factors including achievement goal orientation and regulatory flexibility) [6]. As such, motivation can be viewed as a composite of several variables including, but not limited to: self-efficacy, achievement goal orientation, task value or relevance, and the subjective perception of the task’s demands [10] [11] [12] [13]. Although motivation can be conceptualized as a combination of these constructs, in this paper we focus on achievement goal orientation. More specifically, because achievement goal orientation can identify individual differences for goals in learning or achievement situations, it can be viewed as one of the more important constructs in determining learner success when interacting with ITSs. As some learners are hypothesized to be intrinsically motivated to learn the material, they are often found to be more successful than learners that are more interested in their overall performance. As such, it is important to identify these individual differences to best facilitate, scaffold, and support overall learning with an ITS.

Achievement goals are an important component of motivation in that they offer a purpose or focus for dealing with obstacles encountered in learning situations. Goal orientations guide learners’ behaviors, performance, and affective responses by setting standards by which to assess achievement [14]. Goal orientation is composed of two dimensions: mastery goal orientation and performance goal orientation. Elliot and Murayama’s (2008) framework proposes that learners who implement a mastery goal orientation focus on developing competency and skills, whereas learners who adopt a performance goal orientation focus on outperforming their peers. Each of these dimensions is also composed of an approach and avoidance component. The approach orientation focus on positive normative standards, and the avoidance orientation focus on negative normative standards [10]. This results in four total groups; mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance.

Achievement goals can be used as a framework for understanding learners’ overall motivational states during episodes of complex learning, as they influence learners’ perceptions of the tasks that they must engage to achieve their goals. As such, a learner’s goal orientation can be used to calibrate an ITS to account for the individual differences, which are unique to these different types of learners. For example, mastery-approach learners who are intrinsically motivated to learn will not need the same types of scaffolding as other learners. In fact, some research suggests that too much scaffolding may actually have a negative impact on these learners [15]. Conversely, performance-approach learners may be more open to scaffolding as it helps to improve their performance. However, they may require additional scaffolding to produce the same results as mastery oriented learners. These individual differences may also result in ITSs eliciting different learning-centered affective responses from these groups of learners. This is important to consider, as research has shown that some affective responses can impact a learner’s cognitive, metacognitive, and motivational processes during complex learning [2] [16]. Therefore, the consideration of learner goal orientation may have a significant impact on ITS performance and adaptivity.

2 Method

2.1 Participants and Experimental Design

194 college students (53% female) from three North American universities participated in a 2-day laboratory study. The students ranged in age from 18 to 41 (\(M = 20.46, SD = 2.92\)). Learners received up to $40 for completing the study.
MetaTutor is a hypermedia-based ITS that includes 47 pages of text and static diagrams of the human circulatory system [3]. MetaTutor has now undergone several iterations and has been used as both a tool for research and learning. The system enables the collection of rich trace data on learner SRL use as well as fosters conceptual understanding about the human circulatory system. MetaTutor uses natural language processing to assist learners in understanding challenging content by setting subgoals and advancing the learners toward those goals [4]. Learners are able to select text and diagrams from a table of contents, as well as self-initiate cognitive and metacognitive SRL processes by using the SRL palette of the interface (see Fig. 1).

2.3 Materials, Coding and Scoring

Over the course of both MetaTutor sessions, learners were asked to complete pre- and post-tests, a demographic questionnaire, and several self-report measures, including the Achievement Goal Questionnaire-Revised (AGQ-R; [14]) and Achievement Emotions Questionnaire (AEQ; [17]). The pre- and post-test measures were 30-item multiple-choice tests on the human circulatory system. There were two versions of the test (A and B). The tests were counterbalanced such that if the learners were administered Test A for the pretest they received Test B for the posttest.

The AGQ-R is a 12-item measure that was used to assess learner motivation via their achievement goal orientation [14]. Achievement goals influence learners’ specific goals or plans for a task or obstacle. Achievement goal orientations are separated into four groups: mastery-approach ($\alpha = .84$), mastery-avoidance ($\alpha = .88$), performance-approach ($\alpha = .92$), and performance-avoidance ($\alpha = .94$). The values corresponding to these categories are summed to identify dominant goal orientation.

The AEQ was designed to assess participants’ achievement emotions in learning, class, and testing situations [17]. The questionnaire measured several discrete learning-related trait emotions (i.e., enjoyment ($\alpha = .89$ to .90), hope ($\alpha = .84$ to .89), pride ($\alpha = .84$ to .92), anger ($\alpha = .85$ to .89), anxiety ($\alpha = .89$ to .94), shame ($\alpha = .90$ to .93), hopelessness ($\alpha = .88$ to .94), and boredom ($\alpha = .93$).

A proportional learning gain score was calculated for each learner, using pre- and posttest scores, applying formulas presented by [18] and shown below in Equation 1.

$$\text{Proportional learning gains} = \begin{cases} \frac{\text{post} - \text{pre}}{\text{pre}}, & \text{post} > \text{pre} \\ \frac{\text{post} - \text{pre}}{\text{post} - \text{pre}}, & \text{post} \leq \text{pre} \end{cases}$$

2.5 Experimental Procedure

The study was conducted over two sessions that took place within a 3-day span. Learners were instrumented and calibrated for physiological measures. MetaTutor enables the collection of multichannel SRL trace data throughout the session. Several
self-report measures of emotions and motivation were administered throughout the session, and learners also completed pre and posttests following each session.

3 Results

3.1 Research Question 1: Are there unique clusters of achievement goal oriented learners?

The traditional approach to analyzing achievement goal orientation categorizes learners based on the scale with the highest sum [15]. Thus, a learner with a highest rating of mastery-approach would be assigned to that category even if performance-approach and performance-avoidance were also very highly rated. Intuitively, it makes sense that learners would place along a continuum of ratings across these scales, so why should learners be assigned to one of four categories? Indeed, unnecessary discretization of continuous data has been recognized as a methodological flaw [19]. Therefore, we endeavored to explore finer distinctions in how learners reported achievement goal orientation across these scales.

We applied Expectation-Maximization clustering in WEKA to discover whether patterns existed in learners’ reports of achievement goal orientation [20]. This approach identifies how many clusters and central values (centroids) correspond to each cluster. In order to avoid local maxima, clustering was performed across five random seeds. The best-fit model was selected using maximum log likelihood and was comprised of seven clusters. The mean ratings of each achievement goal orientation scale from each cluster centroid are shown in Table 1.

<table>
<thead>
<tr>
<th>Centroids of Goal Orientation Clusters</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mastery-Approach</td>
<td>13</td>
<td>10</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Mastery-Avoidance</td>
<td>12</td>
<td>9</td>
<td>14</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Performance-Approach</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td>9</td>
<td>12</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Performance-Avoidance</td>
<td>12</td>
<td>10</td>
<td>15</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>14</td>
</tr>
</tbody>
</table>

There were a few notable patterns of achievement goal orientation across the seven clusters. Approximately half of learners rated all items highly (C1 & C3), which would complicate assignment to traditional goal orientation categories. These learners reported that they cared about achieving goals for their own intrinsic benefit and to improve others’ evaluations of their performance. Almost a quarter of learners rated all scales highly except for mastery-avoidance (C2 & C7), thus these learners are

---

1 The range of values learners’ self-reported for the Achievement Goal Orientation scale are as follows: Mastery-Approach (5 to 15), Mastery-Avoidance (3 to 15), Performance-Approach (3 to 15), and Performance-Avoidance (3 to 15).
similar to those above, but had lower ratings of avoiding “learning less” or avoiding an “incomplete understanding” of the material. A small proportion of learners rated approach scales highly (C5), such that they were not as concerned with avoiding poor mastery or performance. Finally, a fifth of all learners reported values that aligned with what we would consider mastery-oriented goal orientation (C4 & C6). These learners rated both mastery-approach and mastery-avoidance disproportionately higher than the performance-oriented scales. Based on these responses, the traditional approach of assigning learners to goal orientation categories (e.g., summing the sub-scales) would have categorized these learners as having a mastery-approach orientation, but it would have also done so for many of the learners from the other clusters. Thus, clustering of achievement goal orientation ratings allowed us to discover those learners who actually rated mastery-orientation above all else, while filtering out learners whose self-reports were more ambiguous.

3.2 Research Question 2: Are there differences in proportional learning gains amongst the achievement goal orientation clusters?

Based on the result of research question 1, we were interested in investigating if learners in clusters C4 and C6 attained greater proportional learning gains than learners in the other five clusters. As such, a Mann Whitney U test was conducted to examine proportional learning gains attained by the two groups (C4 & C6 vs. C1, C2, C3, C5, & C7) of learners. Results indicated that the proportional learning gains attained were significantly greater for learners in the (C4 & C6) group associated with mastery-orientation (\(Mdn = 37.86\)) than for learners in the remaining clusters (\(Mdn = 23.08\)), \(U = 2049.0, p = .004\). These findings indicate that learners in the mastery oriented clusters learned, on average, 52% more than learners in clusters not associated with mastery-orientation. These results align conceptually with achievement goal orientation literature whereby mastery goals are associated with various adaptive learning outcomes including deeper learning strategies and greater persistence in overcoming obstacles, which contribute to greater learning outcomes [10].

3.3 Research Question 3: Are there differences in emotions experienced during learning amongst the two groups of clusters?

Mann Whitney U tests were conducted to compare differences between the two groups (C4 & C6) and (C1, C2, C3, C5, & C7) of learners on self-reported results from the Academic Emotions Questionnaire (AEQ). Results indicated that learners in the (C4 & C6) group associated with mastery-orientation reported significantly lower amounts of shame related to learning (\(Mdn = 20.0\)) than did the remaining clusters (\(Mdn = 28.0\)), \(U = 1707.5, p < .001\). These learners also reported lower amounts of hopelessness (\(Mdn = 20.0\)) and anxiety related to learning (\(Mdn = 30.0\)) than did learners in clusters not associated with mastery-orientation, however at non-significant levels, \(U = 2159.0, p = .08\), and \(U = 2151.0, p = .08\), respectively. These results support the extant literature on achievement goal orientation, whereby mastery goal oriented learners experience less negative academic emotions than performance goal oriented learners [16]. This characteristic is important to consider, as research has shown that negative learning centered emotions impact a learner’s cognitive, metacognitive, and motivational processes during complex learning [2] [16].
4 Implications for Designing Intelligent Tutoring Systems that Adaptively Scaffold and Support Motivation

The results of this study offer insight into the importance of learner achievement goal orientation as one of the core components of motivation, as well as open up new methods for teasing apart learner categorization within the construct. The finding that mastery-oriented learners were able to learn 52% more than learners in other clusters speaks to the great importance of identifying learner profiles. Moreover, as research indicates that mastery and performance-oriented learners experience vastly different academic emotions during learning, it necessitates the unique design of scaffolding that can account for these individual differences. In addition, these findings support our assumptions regarding treating motivation as states that fluctuate over time during learning and therefore need to be addressed throughout learning by providing intelligent, adaptive scaffolding and feedback in order to support learning. For example, future ITSs may seek to consider a composite motivation construct, which can account for greater individual differences as well as make detecting, tracking, and scaffolding for motivation in real-time a reality. By integrating learner self-efficacy, achievement goal orientation, task value or relevance, and the subjective perception of the task’s demands, researchers can more accurately detect learner overall motivation and design adaptive scaffolding to more efficiently account for individual differences. As the focus of this discussion revolves predominantly around value judgments, it is important to note that control is also a very important component of motivation. However we will leave control constructs for a later discussion due to space constraints. A description of several key variables that compose motivation and their relationship to the construct and its measurement are discussed below.

Self-efficacy is the subjective judgment a learner makes about their perceived ability in general or in a specific domain. A learner may be more motivated to learn about concepts that they feel more efficacious with or vice versa. Mastery-oriented learners are likely to have higher general self-efficacy than others. Results of this study which indicate lower levels of shame for this group of learners supports this idea. These learners probably feel less shame as they work hard to overcome challenging task content and perceive learning as an intrinsically valuable pursuit that is achieved through their own effort. Self-efficacy can be determined through prior knowledge assessments as well as by timely prompts during the learning session that ask learners to rate their feelings of efficacy for engaging with informational sources (i.e., Self-Efficacy Inventory in MetaTutor). This information can serve to inform the system of what types of material and what level of difficulty the learner is able to engage with. Next-generation systems that employ adaptive scaffolding may be able to use physiological and contextual signatures to detect and scaffold the learner’s efficacy for the task or particular strategy being used. This advance will require researchers to overcome several major challenges; detecting motivational states in real-time using multi-channel data will pose a significant challenge because fluctuations will occur at several temporal scales (e.g., minute-level analyses for interest based on pupil dilation from eye-tracking data on individual pages to log-file analyses of time spent on individual multimedia content and sequence of navigational patterns across pages of multimedia content) most likely followed by the administration of a self-report measure to ground the trace data inference prior to providing the adaptive scaffolding to the learner.
A second major motivational component is task value, which has been operationally defined as the learner’s subjective perception about the relevance of the task towards achieving their goals. Research has found task value to be predictive of learning performance [11]. Conceivably, achievement goal orientation contributes to task value appraisals in that it impacts the learner’s perception of the task at hand. Perhaps a mastery-oriented learner is more willing to engage with materials that are on the periphery of what they find relevant, whereas performance-oriented individuals will be quicker to dismiss material that is not easily seen as relevant to their goal or useful for outperforming their peers. This leads to mastery-oriented learners engaging with more information over the course of the session than performance-oriented learners, and thus contributing to their increased performance. Task value, like self-efficacy, can be determined through timely prompts, whereby learners are questioned about the relevance and criticality of the tasks that they encounter during the learning session. Understanding the learners’ perception of their tasks will allow for more efficient scaffolding and intervention as well as give researchers a clearer indication of the learners’ thought process while engaging their tasks. Just as for self-efficacy, the same challenges exist in detecting, tracking, modeling, inferring fluctuations in task value and providing adaptive scaffolding with ITSs.

Finally, the learner’s subjective perception of the task’s demands may enable researchers to detect and track motivation in real-time. An integrative approach posited by [13] can be used to describe and measure a learner’s motivation as he or she engages with learning materials [21]. The approach utilizes Obrist’s active-coping hypothesis [22] as well as Brehm’s theory of motivational intensity [23]. This theory posits that the task’s demands are a function of what the learner concludes they must do to successfully accomplish the task. The learner sets an upper limit on the amount of effort they are willing to expend on the task and then attempts to overcome the obstacle. Systolic blood pressure (SBP) has been found to be an efficient measurement of task demand. It is known to increase systematically with task difficulty up to the point where successful performance is perceived to still be possible. However, when the task is seen as irrelevant, or unachievable, SBP declines, [24]. As such, SBP can be measured and used in real-time to trigger timely interventions that can assist learners in overcoming difficult obstacles, as well as determine when learners are becoming unmotivated to engage with the system. Achievement goal orientation impacts this construct in that mastery-oriented learners are known to be more persistent in engaging their tasks. As such, it is likely that mastery-oriented learners would recalibrate their upper limits when they have been unsuccessful in completing a task. Whereas, performance-oriented individuals may be quicker to lose motivation after failing to overcome an obstacle. Identifying these fluctuations in motivation using physiological sensors stands to contribute to our understanding of the dynamics of motivational processes as learners engage in complex learning with ITSs. In conclusion, understanding motivation stands to be a difficult challenge due to major conceptual, methodological, and analytical issues that have severely hampered ITS researchers in addressing learners’ motivation. However, our findings provide evidence that longstanding assumptions can be re-conceptualized as we treat motivation as a critical part of SRL and more specifically treat it as a series of events that can be detected, tracked, modeled with multichannel data (e.g., physiological sensors, eye-tracker, log-files, etc.) and used to design intelligent, adaptive motivation scaffolding. Future research should explore various conceptualizations of motivation such as this, and uti-
lize multichannel trace data to assess, monitor, and scaffold for motivation constructs such as achievement goal orientation and others.

**Acknowledgements**

This study was supported by funding from the National Science Foundation (DRL 1431552). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

**References**


Human Expert Labeling Process (HELP):
Towards a Reliable Higher-Order User State Labeling
by Human Experts

Sinem Aslan, Sinem Emine Mete, Eda Okur, Ece Oktay, Nese Alyuz, Utku Genc,
David Stanhill, Asli Arslan Esme

Intel Corporation
{sinem.aslan, sinem.mete, eda.okur, ece.oktay, nese.alyuz.civici, utku.genc, david.stanhill,
asli.arslan.esme}@intel.com

Abstract. In our longitudinal research, we have been working towards an adaptive learning system automatically detecting student engagement as a higher-order user state in real-time. The labeled data necessary for supervised learning can be obtained through labeling conducted by human experts. Using multiple labelers to label collected data and obtaining agreement among different labelers on same samples of data is critical to train final engagement model accurately. Addressing these challenges, we developed a rigorous labeling process (HELP) specific to educational context with multi-faceted labels and multiple expert labelers. HELP has three distinct stages: (1) Pre-Labeling, including planning, labeler recruitment, training, and evaluation steps; (2) Labeling, involving actual labeling conducted by multiple labelers, and related steps for formative assessment of their performance; and (3) Post-Labeling, generating final labels and agreement measures through processing multiple decisions. In this paper, we outline proposed methods in HELP and describe the developed labeling tool.

Keywords: Labeling, higher-order user understanding, student engagement, ITS

1 Introduction

In our longitudinal research, we have been working towards an adaptive learning system incorporating machine learning and perceptual computing to detect student engagement as a higher-order user state in real-time [1]. To classify engagement level of a student, supervised machine learning is preferred where a large amount of data is needed to be labeled for training engagement model. For traditional object classification problems, labeling process involves coding of scenes indicating presence/absence of a specific object/activity. Such problems are rather objective and can be handled by any trained non-expert. In such contexts, guidelines to train labelers can be very obvious: If you see a car in the picture, label the scene as a picture with a car.

As opposed to such problems, labeling student states in a classroom environment (i.e., in the wild) requires a lot of cognitive processing as there are many variables for labelers to consider before selecting a certain label. For example, a student playing with her hair can signify either a state of boredom or frustration, and this gesture itself cannot be used to decide on a label. On the top of this context-complexity, our research focuses on multi-dimensional student states which we call as higher-order user states.
For student engagement, we incorporate two dimensions of labeling: Behavioral and emotional (i.e., affective). With behavioral labeling, we aim to capture how much a student is into a learning task, whereas with emotional labeling, we try to understand the student’s emotional experience during the learning task. A composite of behavioral and emotional labels would give a holistic picture of the student’s engagement state. By nature, such states are ambiguous and labeling those states are relatively subjective. Thereby, to obtain ground truth labels as accurate as possible, rigorous labeling methods should be implemented.

In this line of research, the majority of studies implemented a labeling process requiring one labeler only - some conducting post facto labeling – labeling after data collection - [2], [3] while some others doing in vivo labeling – labeling during data collection. A well-known example for the latter is BROMP [4], [5], [6]. BROMP is a highly preferred labeling protocol used to record observations of student behaviors and/or emotions in authentic field settings. In BROMP, students are observed in their classrooms one at a time through a round-robin technique by an observer (See the study by Bosch et al. [4] to review BROMP in action). BROMP has main advantages in terms of labeling process such as time (i.e., completing data labeling at the end of a lesson) and resource (i.e., decreasing labeling cost - any instance of a student is labeled by only one observer). However, there are some main challenges unaddressed:

1. Limited chance for revisions: As BROMP requires an observer to complete real-time labeling within a time frame of a class, there is minimal chance for the observer to make any changes backward.
2. Real-time, complex decision making: For a holistic judgment of a student’s state, BROMP suggests both monitoring facial expressions, speech, body posture, gestures, and the student’s interactions with computer. However, from a cognitive-processing perspective, it seems impossible to perform all of these in real-time - without having any option to stop the time and think about a final label to assign.
3. Fragmented labeling experience: BROMP requires a labeler to conduct observations using round-robin technique – spending short amount of time with one student and moving to the next. This results in fragmented labeling experience instead of observing one student throughout the whole time.
4. Limited labels for training a model: Round-robin technique in BROMP results in loss of data and thereby labels. There is a high risk of disregarding important state changes. This signifies a need for continuous labeling.
5. Observer effect: Although having an observer to label student data within a classroom can be advantageous in terms of labelers’ having a more authentic labeling experience, there is an inevitable observer effect associated [7].

To address the above challenges, we developed a set of guidelines enabling a reliable higher-order user state labeling by human experts: Human Expert Labeling Process - HELP. Throughout our longitudinal, 3-year research with three pilots conducted with students in authentic classrooms, we have been consolidating and refining this process [1]. We have been conducting various experiments for improving our labeling methods and tools. In this paper, we outline the details of HELP together with a description of our labeling tool as well as the preliminary experiments and results.
2 HELP

HELP has three distinct stages (See Figure 1): (1) Pre-Labeling - planning, labeler recruitment, training, and evaluation steps; (2) Labeling - actual labeling conducted by multiple labelers along with steps for formative assessment for their performance; and (3) Post-Labeling - generating final labels and agreement measures through processing multiple labelers’ decisions.

![Diagram of HELP stages](image)

**Fig. 1. Overview of HELP**

2.1 Stage 1: Pre-Labeling

**Planning**

a. At this very first stage, it is necessary to design how to train prospective labelers, what sort of materials and tools to use during training, and criteria to select the best subset of labelers to be recruited. Labeler training has two main targets: (1) Train labelers on labeling process, and disqualify those with low agreement levels; and (2) fine tune labeling process according to feedback from labelers to maximize agreement level.

b. Researchers review related literature and leverage their own educational expertise to decide on appropriate labels for labeling of behavioral and emotional states. Based on our research, we decided to use the labels given in the next step (c) taking the circumplex model of affect [8] as a reference.

c. The definitions and examples of each selected label are created based on literature [4], [9], [10], [11], discussions with teachers, and students’ observable behaviors.
Behavioral labels:
- **On-Task:** If the student is active in the learning task (e.g., s/he is watching relevant instructional videos/solving questions, etc.).
- **Off-Task:** If the student is not active in the learning task.
- **Can’t Decide:** If the labeler cannot decide on the behavioral state.
- **Not Available (N/A):** If the data cannot be labeled (e.g., while the student is preparing to leave the class at the end of session).

Emotional labels:
- **Satisfied:** If the student is not having any emotional problems during the learning task. This can include all positive states of the student from being neutral to being excited during the learning task.
- **Bored:** If the student is feeling bored during the learning task.
- **Confused:** If the student is getting confused during the learning task – in some cases this state might include some other negative states such as frustration.
- **Can’t Decide:** If the labeler cannot decide on the emotional state.
- **Not Available (N/A):** If the data cannot be labeled (e.g., while the student is preparing to leave the class at the end of the session).

d. After determining the labels and their operational definitions, a sample chunk of data is selected to be labeled in the labeling practice by researchers. While selecting these data, certain considerations are made:
  i. For behavioral labeling practice, 2 x ~6 min segments are chosen, as representatives of the two major labels (i.e., On-Task, Off-Task), from both assessment and instructional activities.
  ii. For emotional labeling practice, 3 x ~10 min segments are chosen, as representatives of the three major emotional labels (i.e., Satisfied, Bored, Confused), selecting both from assessment and instruction.

  e. The data chunks are labeled by researchers to check validity of the examples and definitions of the given labels prior to sharing them with prospective labelers.

  f. Next, training handouts including the definitions and examples of each label for both behavioral and emotional labeling are prepared. Additionally, labeling tool user manual is prepared to guide labelers during the training session (See Section 3 for the labeling tool).

  g. Lastly, the qualifications of the labelers in terms of their professional background are identified. As we incorporate higher-order user states in our research, we use educational psychologists for making valid labeling of students’ complex states.

Labeler Recruitment, Training, and Evaluation

a. A group of prospective labelers with educational psychology background is invited to take part in the training session.

b. Researchers describe the project and labeling job requirements, explain the definitions of each label by giving and demonstrating some examples from the project pilots. Then, the labeling tool and its functionalities are explained to the prospective labelers. The detailed procedure followed at this stage, for both behavioral and emotional labeling, is described below. Note that these five steps are conducted individually for each prospective labeler and each candidate does labeling for both behavioral and emotional states.
1. Describe setup and context of the collected data.
2. Give (printed) definitions of labels to prospective labelers, and explain these states within the context of 1:1 learning.
3. Demonstrate how to use the labeling tool.
4. Using the tool, present examples for different students’ states. Demonstrate what clues were used (e.g., face, head motion, posture).
5. Discuss the examples and make sure all candidates are on the same page at a high level for definitions. Note that labeler training requires a more descriptive approach rather than prescriptive as higher-order user states are ambiguous in nature. We empower expert labelers to make decisions using their expertise in educational psychology.

c. In the practice part of the training session, specific chunk of data is labeled by individual prospective labelers.

d. The practice part is divided into two rounds. First, prospective labelers label the selected chunk of data for behavioral labeling and then they continue to label the selected data for emotional labeling. After each round, a group discussion is created to get feedback from labelers and provide them guidance to achieve a mutual agreement about labeling specific cases.

e. At the end of this training, researchers analyzed prospective labelers’ data and their labels using a rigorous evaluation procedure. Based on these results, labelers are recruited and assigned to behavioral and emotional labeling tasks separately. At the end, we assign three labelers for behavioral labeling and five labelers for emotional labeling. We assign labelers with top agreement levels to emotional labeling due to relative complexity.

2.2 Stage 2: Labeling

a. Labelers start labeling the scheduled data on a daily basis. For our case, we have them work 2.5 days a week.

b. During the whole labeling process, labelers are monitored regularly to catch any outliers. Towards this end, we created a module within our tool to visually monitor labels across different labelers for the same data. We use this module on a weekly basis to check any discrepancies, together with the overall inter-rater agreement measures calculated over the available data using the Krippendorff’s Alpha [12]. Based on our findings, we give formative feedback to labelers.

c. To monitor and keep track of labelers’ questions during the process, we use a Q&A document to input such questions and how we address them. On a weekly basis, we share the updated document with all labelers. This way, they can see some of the specific questions coming from different labelers and how we address them.

d. At the end of the labeling process, a questionnaire is delivered to understand labelers’ experiences (e.g., strengths and weaknesses of the labeling process, labeling tool, etc.). We use this feedback to improve the next labeling cycle.
2.3 Stage 3: Post-Labeling

a. The inter-rater agreement measures both for behavioral and emotional labelers are calculated using the Krippendorff’s Alpha [12]. In case of any significant outlier, the corresponding labelers are discarded from the subsequent process of deciding on the final labels.

b. After filtering out any significantly inconsistent labelers’ data, it is necessary to define a method to decide on the final labels. In general, when there is a majority among the labelers, the mostly voted label is assigned as the final label. As for all of the labels, the instances with majority on ‘Can’t Decide’ are labeled accordingly. However, the instances of class ‘Can’t Decide’ can easily be extended: If there is a strong disagreement, these instances are labeled as ‘Can’t Decide’.

3 Labeling Tool

In the data collection sessions, students used a content platform which enabled them to conduct instructional activities (e.g., watching instructional videos, reading instructional articles) and do exercises (i.e., assessment activities) on a laptop computer. During these activities, the video of the student with a 3D camera (i.e., Intel® RealSense™ Camera F200), desktop videos, and context and performance logs from the content platform are recorded. Visualization of the tool enables displaying information from different modalities, such as the data collected by the camera and data collected from the content platform’s event logs. Our labeling tool enables an external observer to label the data by assigning pre-defined labels to the labeler-defined session segments (See Figure 2).

Fig. 2. Labeling tool visualized for the behavioral engagement

The collected data are displayed in the form of two video streams: (1) RGB videos and (2) desktop videos of individual students. The tool incorporates playback controllers to facilitate data visualization and labeling. Moreover, different contextual data segments are displayed along the timeline with different colors: blue for instruction and grey for assessment segments. Contextual data such as student ID, session number, question number in exercise segment, or attempt number to solve questions are visualized in the text fields together with date and time of the data collection. The ability to
jump to the next/previous video/article or exercise segment is enabled. To improve labeling experience and increase accuracy of labeling, audio data recorded during the data collection sessions are also integrated into the labeling tool. The tool requires labelers to assign labels to the segments they define based on the state changes. The assigned labels are then visualized along the timeline.

4 Preliminary Experiments and Results

We carried out a number of preliminary experiments towards consolidating and refining procedures of HELP. The experiments aim to answer three major research questions:

1. Is appearance modality (i.e., video-recording of students) on itself enough for labelers to accurately label students’ emotional states or is it necessary to include contextual modality (i.e., desktop videos of students)?
2. For emotional states, is it practical to define a separate state for each quadrant of the circumplex model [8] (i.e., the arousal-valence graph)?
3. Using HELP, can we achieve an acceptable agreement among labelers?

Note that, our list of experiments are not limited to these. Our research is still ongoing with various experiments to improve the proposed process. In the following sections, each experiment will address each research question above.

4.1 Experiment 1: Combination of Modalities

For this experiment, we used six hours of student data (recorded while students were reading articles and solving exercises on a digital platform) and asked three labelers to label the data based on emotional states using two different labeling tools. In the first tool, the labelers labeled the data where the tool included both the students’ recorded videos and their desktop videos. A week after that, the labelers labeled the data using a different version of the tool where they could only see the recorded videos of the students, without their desktop videos.

In this experiment, we aimed to observe the effects of using contextual information (i.e., students’ desktop videos) during the labeling process. We incorporated a qualitative method to understand such effects as we were mostly interested in how labelers made sense of information they had on the tool. Towards this end, we applied “think-aloud-protocol” [13] during the labeling. We recorded the labelers’ speech and videos while using the two different labeling applications (i.e., with and without desktop videos). We conducted thematic analysis on these data [14].

The results show that when there is no desktop videos, the labelers mostly had difficulty to decide between ‘Satisfied’ and ‘Bored’ states. Additionally, the labelers indicated that they benefited from contextual information during the labeling process. The labelers stated having difficulty when labeling without seeing the desktop videos of the students. This implies that providing students’ desktop videos is important for labeling.
4.2 Experiment 2: Selection of Learning-Related Emotional States

We base the foundations of engagement modeling on the work by Woolf et al. [9], where we define engagement as a combination of behavioral and emotional states based on their desirability value (See Table 1). For emotional states, previously we assigned one state for each quadrant of the circumplex model [8] - having four states in total: ‘Excited’, ‘Calm’, ‘Bored’, and ‘Confused’. However, during the initial labeling trials, we had post-interviews with the labelers where the results revealed that the distinction between positive and negative arousal for positive valence states was not clear (e.g., ‘Excited’ vs. ‘Calm’). In addition to this feedback, as illustrated in Table 1, the two positive valence quadrants can be treated in the same way, considering either desirability levels (as in [9]) or engagement states. Based on these findings, we merged the positive valence states (‘Excited’, ‘Calm’) into a single one: ‘Satisfied’.

Table 1. Theoretical Foundations for Engagement Labeling (adapted from [9])

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>On-Task</td>
<td>Highly Motivated/Excited</td>
<td>Highly Desirable</td>
<td>Engaged</td>
</tr>
<tr>
<td>On-Task</td>
<td>Calm/Satisfied</td>
<td>Highly Desirable</td>
<td>Engaged</td>
</tr>
<tr>
<td>On-Task</td>
<td>Confused/Frustrated</td>
<td>Maybe Desirable</td>
<td>Engaged</td>
</tr>
<tr>
<td>Off-Task</td>
<td>Highly Motivated/Excited</td>
<td>Not Desirable</td>
<td>Not Engaged</td>
</tr>
<tr>
<td>Off-Task</td>
<td>Calm/Satisfied</td>
<td>Not Desirable</td>
<td>Not Engaged</td>
</tr>
<tr>
<td>Off-Task</td>
<td>Confused/Frustrated</td>
<td>Not Desirable</td>
<td>Not Engaged</td>
</tr>
<tr>
<td>Off-Task</td>
<td>Bored</td>
<td>Not Desirable</td>
<td>Not Engaged</td>
</tr>
</tbody>
</table>

To reinforce this decision, we performed an experiment, where 10 hours of student data 1 (recorded while the students were watching math videos and solving related exercise questions on a content platform) were labeled by three labelers with two different label sets: One with six states where each quadrant is represented separately (‘Excited’, ‘Calm’, ‘Bored’, ‘Confused’, ‘Can’t Decide’, ‘N/A’); and the other with five states, where positive valence states are merged (‘Satisfied’, ‘Bored’, ‘Confused’, ‘Can’t Decide’, ‘N/A’). When the subset of data was labeled with a single ‘Satisfied’ state, the inter-rater agreement level computed using the Krippendorff’s Alpha was approximately doubled. Although an increase in agreement is expected by switching from six to five states, the substantial improvement indicates that the better labeling was achieved by having one positive valence state.

4.3 Experiment 3: Inter-Rater Agreement Measures

On a dataset of approximately 30 hours collected from 12 students 2 in four one-hour sessions in an authentic classroom (where students were watching math videos and solving related exercise questions on a content platform), we utilized HELP to obtain both behavioral and emotional engagement states of the students. The agreement among

---

1 Four 9th grade students: Three males and one female.
2 12 9th grade students; Nine males and three females.
the labelers is expected to highlight the subjective nature of the task, both for the behavioral and the emotional labeling. Therefore, both types of labeling, we computed the overall agreement: For behavioral, the inter-rater agreement of three labelers over four states; and for emotional, the inter-rater agreement of the five labelers over five states were calculated. We experimented with four different measures, namely Krippendorff’s Alpha, Fleiss’ Kappa, Cohen’s Kappa, Scott’s Pi [15], to investigate whether the choice of metric will affect the results. Agreement measures are summarized in Table 2. As similar results were achieved by different metrics, we utilize Krippendorff’s Alpha in HELP (as given in Section 2.3 (b)) due to its applicability to multiple labelers. The agreement measures reported in Table 2 show that especially for emotional labeling, low-to-moderate agreement is achieved. This indicates that emotional labeling is a subjective task as expected and it is necessary to have multiple number of labelers for each instance and to apply majority voting to obtain final decisions.

Table 2. Inter-rater Agreement Measures for Engagement Labels

<table>
<thead>
<tr>
<th>Engagement</th>
<th>State Count</th>
<th>Labeler Count</th>
<th>Krippendorff’s Alpha</th>
<th>Fleiss’ Kappa</th>
<th>Cohen’s Kappa</th>
<th>Scott’s Pi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral</td>
<td>4</td>
<td>3</td>
<td>0.814</td>
<td>0.835</td>
<td>0.824</td>
<td>0.824</td>
</tr>
<tr>
<td>Emotional</td>
<td>5</td>
<td>5</td>
<td>0.542</td>
<td>0.559</td>
<td>0.544</td>
<td>0.545</td>
</tr>
</tbody>
</table>

5 Conclusion

For an improved performance in supervised strategies utilized, we developed a rigorous labeling process specific to educational context with multi-faceted labels and multiple expert labelers. In this paper, we outlined the details of this process along with a labeling tool developed as a part of our longitudinal study. As our research is on-going, from a design-based research perspective, we will continue refining this process and the labeling tool towards a more reliable higher-order user state labeling by human experts.

References


Modeling Visual Attention of Students Playing an Educational Game for Physics

Celesmae T. Vicentea, Ma. Mercedes T. Rodrigoa, Hong-Fa Hob, Guan-An Chenb

aDepartment of Information Systems and Computer Science, Ateneo de Manila University
bDepartment of Electrical Engineering, National Taiwan Normal University

celesmae2010@gmail.com, mrodrigo@ateneo.edu,
jackoho@ntnu.edu.tw, crackereidolon@gmail.com

Abstract. This study models the visual attention of students as they think of a solution to a problem within an educational game for Physics. Participants were given time to view a hint followed by a static image of the game problem. Upon viewing the problem, students were instructed to think of a solution using the hint while an eye tracker recorded eye movement data. After viewing the problem, participants played the actual game level. The game awarded gold, silver, or no badges to participants depending upon their performance. When analyzing the relationship between the eye movement and performance, the findings were that participants who earned gold, silver, and no badges had different orders of fixating on the regions of interest while thinking of a solution to the problem and participants who had better performance fixated earlier on the regions where the solutions should have been drawn.

Keywords: Attention, Physics Playground, Eye Tracking

1 Introduction

Attention refers to concentration given to one phenomenon, to the exclusion of others [10]. It requires tuning out of other stimuli so that a person can apply himself or herself to the phenomenon of interest [8]. The focus of this paper is attention in the context of learning. Research has shown that student performance levels are associated with endurance or attention span [4]. Fluency or accuracy of learner performance in significant durations has been linked with higher or sustained attention [4]. In this study, we examine the relationship between attention and performance among students using Physics Playground (PP), an educational game for designed to help secondary students understand Newtonian Physics [14]. PP helps students better understand balance, mass, conservation and transfer of momentum, gravity and potential kinetic energy [13]. PP has 80 levels. The goal in each level is for students to guide a green ball to a red balloon by drawing machines such as the inclined plane or ramps, levers, pendulums, and springboards. PP awards badges to the players depending upon the number of objects the player used. A gold badge is given when players solve a problem at or below a par number of objects for the problem. Silver badges, on the other hand, are awarded when the player solves the problem with a greater than par number of objects. All player actions are recorded in the log files [13].

© Springer-Verlag Berlin Heidelberg 2011
PP’s impact on student achievement is varied. Studies in the US [3] showed that PP is able to augment students’ understanding of qualitative physics. In the Philippines, however, PP did not result in any learning gains [1][2]. Why is it that so? This study investigates the role of attention (or lack thereof) among Philippine students playing PP using eye movement.

Just and Carpenter provide some of the earliest studies on attention and eye-movements. They coined the “eye-mind” hypothesis which pertains to the notion that eye-movement is a “on top of the stack” indicator of cognitive process and that it can give a trace of where attention is directed [9]. Because of this, researchers have developed means of quantifying attention. One of which is the chosen method in conducting this study, eye-tracking. Eye-tracking is a method that quantifies eye movements relative to the current line of sight of a participant in a stimulus at a given time. When eye movements are quantified, they are broken down into two major metrics, saccades and fixations. Saccades are sudden changes of eye gazes between points of fixation [11]. While fixation is maintaining of eye gaze at a certain position indicating a person’s intention of what to interact with [16], what current task a person is working on [15] and is a proxy indicator of attention[5]. Metrics on fixation and gaze durations have been used in attempt to find answers to research questions of this study. This paper attempts to determine the relationship between eye gaze and student performance.

2 Methodology

2.1 Participants

A total of 30 high school students were recruited from different parts of the Philippines for this experiment. There were 19 male and 11 female participants, 10 of whom were from Luzon, 11 from Visayas, and 9 are from Mindanao. The mean average of participants was 15.5 years and participants were in grade levels 9 to 11. Participants typically spent 2.9 hours a day playing video games and 3.6 hours a day watching television. Participants’ Physics grades’ were 87.9 out of 100 on average. Participants were asked to play the tutorial portion of PP to become acquainted with the game mechanics. After the tutorial, participants viewed static images of a hint and the problem to think of a solution while the eye-tracker was recording their eye movements. Details on this will be discussed in the next section. Participants then played the PP level they viewed. Participants were given a maximum of 3 minutes to solve each level. Finally, participants took a post-test that was isomorphic to the pre-test. Comparison of pre-test and post-test scores showed no improvement after using the PP.

2.2 Material

There were two types of stimuli used in this experiment. The first type (Figure 1, left) contained a hint regarding the simple machine that was most relevant to the solution.
The second type (Figure 1, right) was the static image of the pre-selected game level problem from PP named Scale. Participants were instructed to look at the PP problem while they were thinking of how to bring the green ball to the red balloon. While they were doing this, the eye-tracker recorded their eye-movements. The regions of interest (ROI) of each stimuli were later defined according to the types described below:

1. **Hint** - Simple machine hint that was most relevant to the solution
2. **Instruction** - Instruction about what keys to press to go to the next picture
3. **Starting Point** - Initial position of the green ball
4. **Target** - Location of the red balloon
5. **Solution Space** - Where the simple machine solution is ideally drawn
6. **Decision Factor** - Objects in the stimulus that implicitly contribute to the decisions participants make in solving the problem
7. **Travel Path** - The green ball passes through this area to reach the red balloon
8. **Existing Objects** - Objects that are not explicitly essential in solving the problem

![Fig. 1. Stimulus 1 on the left is the Lever Hint for Stimulus 2 PP Scale level on the right.](image)

There were 2 types of ROIs in the hint, and 6 types of ROIs in the PP level. On the Lever hint, the ROIs defined are: 1) R1 – Hint, 2) R2 – Instructions. The static image for PP game level Scale has the following ROIs: 1) R1 – Starting Point, 2) R2 – Target/Solution Space, 3) R3 – Solution Space, 4) R4 – Decision Factor and 5) R5 – Travel Path. The regions that were not defined among these categories fell under the “others” classification. These defined ROIs are the essential points of investigation and comparison during the analysis. The ROI definitions on the PP game level were assigned on the basis of where solutions were drawn. Figure 2 shows two ways of solving the problem. Objects can be drawn on R3 or R2 to offset the weight of the Scale and bring the green ball to red balloon.

![Fig. 2. Lever solutions to bring the green ball and to the red balloon.](image)
2.3 Instrument

The eye movement data were recorded using the EyeNTNU-120 eye tracker that has been used to study visual attention on reading process on integrated circuits in [6] and on how women perceive handbags in [7]. Figure 3 shows the set-up during the data gathering. Participants were asked to place their chins on the chin-rest while the eye-camera was directed at one of the participant’s eyes. As the participant viewed the onscreen stimuli, the eye-tracker recorded and mapped participants’ eye movements in regions. This device had a sampling rate of 120 Hz and an error rate of less than 0.3 degrees given that the participants were less than 60 centimeters away from the computer screen. The four vital metric variables provided by the system and have been the basis for data analysis are the following:

- Total Contact Time (TCT) – total time in milliseconds a participant gazed on ROI.
- Number of Fixations (NOF) – the number of times the participant fixated on ROI.
- Duration of First Fixation (DFF) – the total time in milliseconds that the first fixation on ROI lasted, and
- Latency of First Fixation (LFF) – the time when the first fixation on ROI occurred.

![Figure 3. EyeNTNU-120 Camera and Chin-rest Set-up](image)

User eye gaze data and user action logs from PP were later synchronized in order to investigate the relationship between eye gaze data with user performance based on badges earned.

3 Results and Discussions

Out of 30 participants, 20 solved the problem. Six participants earned gold badges and the other 14 received silver badges. Ten participants did not earn any badges. The succeeding sections discuss the analysis of the metric values of the ROIs for both
Stimulus 1 and Stimulus 2 and the consolidated ROI grouped by badges earned. Percentages were used to determine the rate of TCT and NOF to normalize data.

3.1 Total Contact Time (TCT)

**Stimulus 1 – Lever Hint for PP Game Level Scale.**

As seen in Table 1, all participants spent more time gazing at R2 - Instructions than in R1 - Hint. For R1, gold badge earners had the highest percentage of TCT (22.71%) followed by those who did not earn any badge (12.6%) and the silver badge earners (7.1%). There is a statistically significant difference on the TCT of R1 among the groups based on one-way ANOVA (F(2,27)=6.7, p = 0.004). A Tukey post-hoc test showed that TCT of gold badge earners were significantly higher than those who earned silver badge (7.1%±6.9%, p = 0.003) and marginally higher than those who did not earn any badge (12.6%±8.4%, p=0.083). For R2 the instructions, silver badge earners had the highest TCT (87.9%), followed by those who did not earn any badge (82.8%) and then the gold badge earners (72.8%). There is a statistically significant difference on the TCT of R2 among the groups based on one-way ANOVA F(2,27)=5.6,p=0.010. Gold badge earners had significantly lower TCT on the instructions than those who earned silver badge (8.8%±7.1%, p=0.007).

<table>
<thead>
<tr>
<th>BADGE</th>
<th>Hint</th>
<th>Instructions</th>
<th>ORDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>2 (22.7%)</td>
<td>1 (72.8%)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td>SILVER</td>
<td>2 (7.1%)</td>
<td>1 (87.9%)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td>NO BADGE</td>
<td>2 (12.6%)</td>
<td>1 (82.8%)</td>
<td>R2 &gt; R1</td>
</tr>
</tbody>
</table>

**Table 1.** TCT percentage per ROI on Stimulus 1 – Hint for PP Game Level Scale

**Stimulus 2 – PP Game Level Scale.**

The order of ROI from the greatest to the least TCT grouped by badge earned in Table 2 shows that Gold badge earners spent the highest time in R2, R4, and R3 which have ROI types Target/Solution Space, Decision Factor and Solution space. Silver badge earners and those who did not earn any badge on the other hand spent less time looking at the top three ROI types than gold badge earners. Results also show that the Gold badge earners had the highest TCT in R2 (solution space) (28.51%) followed by the silver badge earners (16.02%) and those who did not earn any badge (8.82%).

<table>
<thead>
<tr>
<th>BADGE</th>
<th>Starting Point</th>
<th>Target/Solution Space</th>
<th>Solution Space</th>
<th>Decision Factor</th>
<th>Travel Path</th>
<th>ORDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>5 (11.2%)</td>
<td>1 (28%)</td>
<td>3 (14.3%)</td>
<td>2 (19.5%)</td>
<td>4 (13.4%)</td>
<td>R2 &gt; R4 &gt; R3 &gt; R5 &gt; R1</td>
</tr>
<tr>
<td>SILVER</td>
<td>1</td>
<td>3 (5%)</td>
<td>5 (4%)</td>
<td>2 (2%)</td>
<td></td>
<td>R1 &gt; R5 &gt; R2 &gt;</td>
</tr>
</tbody>
</table>

**Table 2.** Percentage and Order of TCT from the Greatest to the Least Grouped by Badge
### Consolidated ROI

R1 of Stimulus 1 and R1 to R5 of Stimulus 2 were consolidated into a single ROI. The TCT of gold, silver, and no badge participants were averaged and compared. Gold, no badge, and silver badge earners spent 18.3%, 15%, and 14.5% of their TCT looking at this consolidated region respectively. The gold badge earners were more engaged in the regions identified to be crucial in accessing information in solving a problem. There was a statistically significant difference between groups as determined by one-way ANOVA (F(2,27)=4.307, p= 0.024). A Tukey post-hoc test revealed that the consolidated TCT of gold badge earners was statistically higher than those who earned silver badge (14.5%±2.6%, p=0.021) and marginally higher than those who did not earn any badge (15.0%±3.0%, p=0.066). There was no statistically significant difference between the gold badge earners and silver badge earners (p=0.893).

### 3.2 Number of Fixations (NOF)

**Stimulus 1 – Lever Hint for PP Game Level Scale.**

Table 3 shows the percentages of NOF per ROI in Stimulus 1. Gold badge earners had the highest NOF percentage on R1 at 20% followed by those who did not earn any badge at 13.1% and finally by silver badge earners at 7.2%. There is a statistically significant difference between groups determined using by one-way ANOVA (F(2,27)=5.367, p=0.011). Gold badge earners had significantly more NOF on R1 than silver badge earners (7.2%±6.3%, p=0.009).

<table>
<thead>
<tr>
<th>BADGE</th>
<th>Hint</th>
<th>Instructions</th>
<th>ORDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>2 (20.0%)</td>
<td>1 (76.0%)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td>SILVER</td>
<td>2 (7.2%)</td>
<td>1 (88.1%)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td>NO BADGE</td>
<td>2 (13.1%)</td>
<td>1 (83.5%)</td>
<td>R2 &gt; R1</td>
</tr>
</tbody>
</table>

**Table 3. NOF Percentage on Stimulus 1 – Hint for PP Game Level Scale**

**Stimulus 2 – PP Game Level Scale.**

Table 4 shows the NOF percentage per ROI in stimuli 2. Ranking NOF percentages from the greatest to the least show that silver badge earners and those who did not earn any badge had the same order. Both groups had the highest fixation on ROIs R5, then R1, then R4. ROIs types R2 and R3 spaces had the least NOF percentage. While gold badge earners on the other hand, had the highest NOF percentage on ROIs R2, R4, and R3 and with the least NOF on the R5 and R1. This is indicative that gold badge earners had higher attention on ROI types Target, Solution Space, and Decision
Factors. While both silver badge and no badge earners had the least attention on these ROI types. There is a marginal significant difference between groups in R2-target location of red balloon, as determined by one-way ANOVA (F(2,27)=2.782, p=0.08). A post-hoc Tukey test showed that gold badge earners have marginally higher NOF than those who did not earn any badge (8.8%±8.8%, p=0.066).

<table>
<thead>
<tr>
<th>Badge</th>
<th>Starting Point</th>
<th>Target/Solution Space</th>
<th>Decision Factor</th>
<th>Travel Path</th>
<th>Order (from greatest to least NOF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>5 (11.3%)</td>
<td>3 (14.3%)</td>
<td>2 (16.3%)</td>
<td>4 (14.1%)</td>
<td>R2 &gt; R4 &gt; R3 &gt; R5 &gt; R1</td>
</tr>
<tr>
<td>SILVER</td>
<td>2 (18.1%)</td>
<td>5 (12.7%)</td>
<td>3 (15.0%)</td>
<td>1 (18.8%)</td>
<td>R5 &gt; R1 &gt; R4 &gt; R2 &gt; R3</td>
</tr>
<tr>
<td>NO BADGE</td>
<td>2 (19.4%)</td>
<td>4 (8.8%)</td>
<td>3 (17.5%)</td>
<td>1 (26.3%)</td>
<td>R5 &gt; R1 &gt; R4 &gt; R2 &gt; R3</td>
</tr>
</tbody>
</table>

Table 4. NOF percentage per ROI on Stimulus 2 – PP Game Level Scale.

Consolidated ROI
R1 of Stimulus 1 and R1 to R5 of Stimulus 2 were consolidated into a single ROI. The NOF of gold, silver, and no badge participants were averaged and compared. Gold, no badge, and silver badge earners spent 17.4%, 15%, and 14.5% of their NOF looking at this super region respectively. There was a marginal significant difference between groups as determined by one-way ANOVA (F(2,27)=2.896, p=0.073). A Tukey post-hoc test revealed that the gold badge earners had marginally significant higher NOF than those who earned silver badge (14.5%±2.4%, p=0.061). There is no significant difference between the NOF of those who did not earn any badge and those who earned gold badge (p=0.180) and those who earned silver badge (p=0.858).

3.3 Latency of First Fixation (LFF) and Duration of First Fixation (DFF)

Figure 4 shows that all participants parsed the ROIs in Stimulus 1 in the same manner. That is, the first fixation was R2 (Instructions), followed by R1 (Hint). On average, Gold badge earners fixated on R1 at the 9.31 second mark, then followed by those who did not earn any badge at the 10.04 second mark and then by silver badge earners at the 16.51 second mark. The DFF orders on the other hand were the same for silver badge earners and those who did not earn any badge where in DFF on R2 was greater than R1. In addition, while the Gold badge earners had the earliest LFF on R1, they also fixated on this region the longest. This indicates that they fixated earlier on the hint and looked at it the longest compared to other participants. Table 5 shows the LFF (rank according to time access, from the earliest LFF) and DFF (rank according to length of time first fixation happened, from the greatest DFF) results for stimulus 1.
Fig. 4. LFF for Stimulus 1
Fig. 5. LFF for Stimulus 2 – Gold Badge Earners
Fig. 6. LFF for Stimulus 2 – Silver Badge
Fig. 7. LFF for Stimulus 2 – No Badge

<table>
<thead>
<tr>
<th>BADGE</th>
<th>Metric</th>
<th>Hint</th>
<th>Instruction</th>
<th>ORDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>LFF</td>
<td>2 (9.3s)</td>
<td>1 (0.3s)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td></td>
<td>DFF</td>
<td>1 (84.5ms)</td>
<td>2 (46.8ms)</td>
<td>R1 &gt; R2</td>
</tr>
<tr>
<td>SILVER</td>
<td>LFF</td>
<td>2 (16.5s)</td>
<td>1 (0.3s)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td></td>
<td>DFF</td>
<td>2 (41.45ms)</td>
<td>1 (43.4 ms)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td>NO BADGE</td>
<td>LFF</td>
<td>2 (10.0s)</td>
<td>1 (0.5s)</td>
<td>R2 &gt; R1</td>
</tr>
<tr>
<td></td>
<td>DFF</td>
<td>2 (38.4ms)</td>
<td>1 (39.9 ms)</td>
<td>R2 &gt; R1</td>
</tr>
</tbody>
</table>

Table 5. LFF and DFF Ranking and Time Averages for Stimulus 1 – Lever Hint

Figures 5 to 7 show the LFF on Stimulus 2 of participants that earned Gold badges, Silver badges, and no badges. Table 6 shows that all groups of participants have different orders LFF and DFFs for Stimulus 2. In addition, Results of LFF showed that Gold badge earners accessed the ROIs that are solution spaces earlier, at 4.80 second mark, whereas those who did not earn any badge at the same ROI at the 8.15 second mark. Interestingly, Gold badge earners accessed the solution space earlier but spent the least time on these spaces themselves. This implies that these participants arrived at the solutions earlier and faster than those who did not earn any badge.

<table>
<thead>
<tr>
<th>Badge</th>
<th>Metric</th>
<th>Starting Point</th>
<th>Target /Solution Space</th>
<th>Solution Space</th>
<th>Decision Factor</th>
<th>Travel Path</th>
<th>ORDER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
<td>R4</td>
<td>R5</td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td>LFF</td>
<td>1</td>
<td>0.8s</td>
<td>2</td>
<td>3.1s</td>
<td>5</td>
<td>6.5s</td>
</tr>
<tr>
<td></td>
<td>DFF</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Silver</td>
<td>LFF</td>
<td>52.5ms</td>
<td>40.0ms</td>
<td>26.3ms</td>
<td>71.2ms</td>
<td>44.0ms</td>
<td>R2 &gt; R3</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>R4 &gt; R1 &gt; R5 &gt; R2 &gt; R3</td>
<td></td>
</tr>
<tr>
<td>DFF</td>
<td>2</td>
<td>2.6s</td>
<td>4.7s</td>
<td>6.7s</td>
<td>2.5s</td>
<td>0.6s</td>
<td>R3 &gt; R1 &gt; R2 &gt; R5 &gt; R4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No Badge</th>
<th>LFF</th>
<th>52.1ms</th>
<th>45.3ms</th>
<th>52.1ms</th>
<th>29.9ms</th>
<th>44.0ms</th>
<th>R2 &gt; R3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2.9s</td>
<td>6.7s</td>
<td>9.7s</td>
<td>4.5s</td>
<td>1.8s</td>
<td>R1 &gt; R5 &gt; R4 &gt; R2 &gt; R3</td>
</tr>
<tr>
<td>DFF</td>
<td>5</td>
<td>39.1ms</td>
<td>48.8ms</td>
<td>40.4ms</td>
<td>48.1ms</td>
<td>52.8ms</td>
<td>R3 &gt; R1</td>
</tr>
</tbody>
</table>

Table 6. LFF and DFF Ranking and Time Averages for Stimulus 2 – PP Game Level Scale

4 Conclusion and Future Work

Results show that students that had good performance paid more attention on the regions essential for solving the physics problem while they were still thinking of a solution. Time sequence in parsing the problem also showed that participants that performed better assessed the problem first by looking at the initial position of the green ball and then looking at the region of red balloon target and the solution space. Then they looked at objects that are contributory to solving the problem and then looked for an alternate solution. This is indicative that participants that had a good performance assessed what needs to be done to solve the problem and had a goal in mind on the first parsing through the regions of the interest. The participants that had a bad performance did not have the same sequence of parsing. In addition, participants who solved the problem accessed the solution spaces almost half the time earlier than those who did not earn any badge. This is indicative that students that earned badges had thought of solutions faster and earlier than those who did not solve the problem. Investigation on other stimuli in PP with different difficulty levels will be done in the future in order to know if a trend on attention levels can be discovered.

Acknowledgement

We thank the Ateneo de Manila University, the Ateneo Laboratory for the Learning Sciences, the National Taiwan Normal University (NTNU), the College of Technology and Engineering of NTNU, Dean Kuan-Chiao Yu, Professor Chung-Yen Su, Juan Miguel Andres, Ryan S. Baker, Valerie Shute, Matthew Ventura, Matthew Small, Jaclyn Ocumpaugh, Jessica Sugay, Juliana Alexandra Andres, Michelle Banawan, Yancy Vance Paredes, and Nicko Rigo Caluya.

We also thank the Department of Science and Technology’s (DOST) Philippine Council for Industry, Energy, and Emerging Technology Research and Development for the grant entitled “Stealth assessment of student conscientiousness, cognitive-affective states, and learning using an educational game for Physics” and also the DOST-Engineering Research and Development Technology’s Sandwich Program
Grant for Internship at National Taiwan Normal University that made this research project possible.

References

The coupling between metacognition and emotions during STEM learning with advanced learning technologies:
A critical analysis and implications for future research

Roger Azevedo,1 Nicholas V. Mudrick,1 Michelle Taub,1 Seth A. Martin,1 Franz Wortha,2 and Garrett C. Millar1

1North Carolina State University, Department of Psychology, Raleigh, NC, USA
2Technische Universität Dresden, Department of Psychology, Dresden, Germany
{razeved@ncsu.edu}

Abstract. Metacognition and emotions play a critical role in learners’ ability to monitor and regulate their learning about science, technology, engineering, and mathematics (STEM) content while using advanced learning technologies (ALTs; e.g., intelligent tutoring systems). In this paper we focus on: (1) presenting a succinct review of the assumptions, strengths, and weaknesses of two leading models of metacognition and emotions typically not adopted by ITS researchers; (2) presenting and critiquing Azevedo and colleagues’ extension of the information processing theory of SRL by articulating the assumptions as well describing the advantages and weaknesses of including macro-level, micro-level, and valence to metacognitive processes; and (3) proposing future directions and presenting instructional implications for the design of metacognitive and affect-sensitive ITSs for STEM.

Keywords. Metacognition, emotions, affect, models, processes, intelligent tutoring systems.

Interplay between Metacognition and Emotions in Advanced Learning Technologies

Metacognition and emotions play a critical role in learners’ ability to monitor and regulate their learning about STEM content while using ALTs (e.g., ITSs, adaptive hypermedia, simulations, serious games). These processes unfold in real time as learners dynamically monitor and attempt to regulate them in order to enhance their learning of complex STEM instructional materials. While the literature is inundated with frameworks, models, and theories that focus on cognition, emotions, and cognition with emotions, there is no comprehensive model that binds emotions with metacognition. We argue that the coupling of metacognition and emotions is needed to fully comprehend learners’ self-regulated learning (SRL), including when they use ALTs for STEM learning. For example, negative emotional responses (e.g., frustration) following a pedagogical agent’s (PA’s) prompts and scaffolding could be based on miscalibrated metacognitive judgments from previous learner-PA interactions and the learner’s perceived utility of the scaffolding; a learner’s prolonged negative emotional state (or transition between negative emotional states) when inspecting STEM content could be triggered by a failure to correctly judge the content’s relevance in relation to the current learning goal (i.e., misappropriated stimulus evaluation check);
a learner may also experience prolonged frustration when attempts to use an optimal learning strategy continuously leads to suboptimal results; and so forth. These are just a few examples that highlight the need to develop a theoretical framework that binds metacognition and emotions. As such, our paper will provide a succinct review of two leading metacognition and emotions theories of Winne & Hadwin [1,2] and Scherer [3] seldom used by intelligent tutoring system (ITS) researchers. In our opinion, these theories are critical to advancing our conceptual and theoretical bases for ITS research as well as having implications for designing metacognitive and affective-sensitive ITSs. More specifically, our paper will focus on: (1) presenting a succinct review of the assumptions, strengths, and weaknesses of two leading models and theories of metacognition [1,2] and emotions [3] typically not adopted by ITS researchers; (2) presenting and critiquing Azevedo and colleagues’ extension of the information processing theory of SRL by articulating the assumptions as well describing the advantages and weaknesses in including macro-level, micro-level, and valence to meta-cognitive processes; and (3) proposing future directions and presenting instructional implications for the design of metacognitive and affect-sensitive ITSs for STEM.

Background, Theoretical Framework, and Related Work

SRL is critical to learning, problem solving, reasoning, and understanding complex topics and domains when using ALTs such as ITSs [4,5]. However, several major issues continue to impede learners’ ability to excel academically when using ALTs to improve the education and training needs of learners of all ages [6]. First, learners are not being adequately prepared to meet the needs of 21st century STEM jobs. Deficiencies include critical thinking, communication, and collaboration skills. Second, success in STEM careers requires the deployment of cognitive, affective, metacognitive, and motivational (CAMM) regulatory processes across complex STEM topics, which students lack training in and frequently fail to deploy. Third, the use of ALTs to enhance students’ learning and problem-solving skills in STEM fields, for the most part, has not been guided by theories of how people learn [6-10].

Research suggests students rely heavily on ineffective strategies, such as copying information, instead of effective strategies, such as summarizing content in their own words, hypothesizing, and making inferences [11]. In addition, they fail to accurately monitor their use of learning strategies, their emerging understanding of the content, and the adequacy and relevance of multiple representations of information afforded by the learning environment; they also fail to relate the informational sources with prior knowledge. Motivationally, they show little interest in science, do not value the task, misattribute their success to external factors, and lack self-efficacy in their ability to use effective strategies [12].1 Emotionally, learners frequently show prolonged frustration, boredom, and confusion, which they are ineffective at regulating to re-engage in learning [13-16]. Finally, despite results on the relevance and importance of CAMM processes, there is a strong need for systematic research focusing on all four

1 We acknowledge the role of motivation, but it is beyond the scope of this paper to cover it here.
CAMM processes simultaneously. For example, recent efforts have focused on cognition and metacognition [5, 17-19], cognition and emotions [20], affect and engagement [15], and motivation and emotions [21], but none have explicitly addressed the conceptual and theoretical link between metacognition and emotions.

We argue this is a major flaw that needs to be addressed by interdisciplinary researchers working in the area of ALTs. More specifically, we argue there is a need to: (1) conceptually and theoretically link assumptions, levels of description, and explanations between metacognition and emotions given their respective fundamental assumptions about monitoring and appraisals; (2) provide an initial process-oriented model/framework between metacognitive processes (including judgments) and appraisal mechanisms (e.g., stimulus-evaluation checks) and their respective control processes (including cognitive processes and emotion regulation strategies); (3) employ trace methodologies to detect, track, and model temporally unfolding metacognitive and appraisal mechanisms during learning with ALTs; and (4) make inferences about the processes in (3) to understand fundamental mechanisms and then embody them in ALTs to provide real-time individualized metacognitive and emotions support to learners.

Overview of the Assumptions of the Information Processing Theory of Self-Regulated Learning

Winne and Hadwin's [1,2] information processing theory (IPT) model of SRL is the leading process-oriented model currently used by interdisciplinary researchers studying SRL-related issues and designing ALTs [5,6,22-24]. The IPT model of SRL [1,2] has several theoretical assumptions regarding the use of SRL processes during each phase of learning. First, there are four distinct cycles (e.g., task definition, goal setting, studying tactics, and adaptation), although they might not be mutually exclusive, such that if students are engaging in studying tactics (e.g., reading content), they can simultaneously adapt these tactics to ones that will be effective for learning, thus also adapting the plans set in the second stage (goal setting and planning). For example, if students are reading a page and determine that it is not relevant to their current sub-goal, the students should change the page and thus adapt their plans and studying tactics. This implies that the students are engaging in phases two, three, and four all at the same time, not independently.

However, this model assumes students are aware of these different stages and processes involved in information processing prior to beginning the task. In fact, the model postulates that monitoring and control are the hubs of SRL. Therefore, students should already know they should set subgoals, plan to use the appropriate cognitive and metacognitive SRL strategies, and know how to adapt the use of these strategies as well. Finally, the IPT model posits that throughout each stage of learning, students are constantly monitoring and controlling how their learning is unfolding such that they are in control of the learning processes they are using, and they are monitoring how effective these processes are in contributing to learning, information processing, and thus task completion.
Macro- and Micro-Level Information Processing Theory Processes

Greene and Azevedo [19] developed a model of SRL by expanding on Winne and Hadwin's [1,2] IPT model and Pintrich's [25] four-phase SRL model. Greene and Azevedo's [19] model expands on these original frameworks by proposing specific micro-level valenced SRL processes. The five macro-level processes introduced in their model are listed as planning (e.g., activating prior knowledge, setting subgoals), monitoring (e.g., feelings of knowing, content evaluation, self-questioning, judgments of learning, monitoring progress towards goals), strategy use (e.g., drawing, coordinating informational sources, knowledge elaboration, maintenance rehearsal, hypothesizing, making inferences), handling of task difficulty and demands (e.g., help-seeking behavior), and interest activities (e.g., interest in content domain [19]). Additionally, 35 specific micro-level SRL processes discovered through extensive research using concurrent think-aloud methods have been categorized within these macro-level SRL processes [19]. Below we describe three specific metacognitive monitoring processes that are particularly relevant to different types of learning with ALTs.

Three distinct yet critical metacognitive monitoring processes include feeling of knowing (FOK), judgment of learning (JOL), and content evaluation (CE). For example, FOK is a micro-level SRL process within the macro-level domain of monitoring [7,26]. This micro-level process consists of participants' realization that they have pre-existing familiarity with the content presented within the learning environment. This typically involves the awareness that the material currently being read or inspected during learning with an ALT has previously been learned. This monitoring process involves activation of previous knowledge stored in long-term memory and potential activation of this knowledge during learning with the ALT. By contrast, as a metacognitive judgment, JOL occurs when learners raise awareness as to if they do (+) or do not (−) comprehend or understand something that has just been read or inspected. It is closely related to FOK considering both processes monitor correspondence between a learner's domain knowledge (i.e., cognitive condition) and the learning resources (i.e., task condition). Another example is CE, whereby a learner compares current learning subgoals vis-à-vis the multiple representations currently being read or inspected and judges their appropriateness (+) or inappropriateness (−) to the current subgoal. Therefore, Azevedo, Greene, and colleagues focus on the macro-level, micro-level, and valence of metacognitive processes that extend and address some of the challenges found in Winne and Hadwin’s model.

The integration of Winne and Hadwin's IPT model with Azevedo, Greene, and colleagues' model, based on the hierarchical nature of macro-level, micro-level, and valence of metacognitive judgments, offers several conceptual, theoretical, methodological, analytical, and design issues. Adding micro-level processes and valence allows for detailed feedback mechanisms to describe the temporally unfolding metacognitive processes (in real time) and the associated adaptive control strategies, based on the valence. For example, while monitoring the relevance of diagrams to one's goals, the learner judges they are not relevant (i.e., CE−) and subsequently switch to the table of contents searching for potentially relevant multiple representations of information available in the ALT. However, there is a possibility that the learner
might have persisted inspecting the diagrams despite their lack of relevance to the current learning goal. In the next section we present Scherer’s component process model [3] by describing its assumptions and then presenting the strengths and weaknesses of the model.

Scherer’s Component Process Model

Scherer’s component process model (CPM) [3,27,28] is a contemporary appraisal theory of emotions based on the conceptualization of emotions as processes rather than states. In its current form the CPM is one of the most comprehensive models that aims to describe emotions as dynamic processes that unfold over time, including interactions among multiple components and processing on several levels [3]. According to the CPM, emotions arise from different appraisals of a relevant event. The appraisal sequence includes the following four steps: (1) relevance, (2) implications, (3) coping, and (4) normative significance. The appraisals steps consist of discrete sets of stimulus evaluation checks (SECs) that define different appraisal outcomes. For example, evaluation of the relevance of an event includes assessments of its novelty, intrinsic pleasantness, and goal conduciveness. The appraisals are postulated to occur in a fixed order, but Scherer [3] highlights that the emotional process is recursive by nature, which means that some appraisal outcomes might be revisited and updated during an emotion episode and that some of those processes even occur simultaneously. Furthermore, the appraisal process both influences and is influenced by other components of the emotion process including the subjective feeling component, motor expressions, action tendencies, and physiological changes, as well as other psychological constructs and processes such as attention, memory, or motivation. These interactions not only shape the ongoing emotional process but can also feed forward into future emotion processes. Ultimately, emotions then lead to motivational changes, and prepare action readiness and action tendencies. However, they are not solely sufficient causes for the execution of actions; rather, the execution of actions is multiply determined depending on other factors (e.g., volition or cognitive control [3]).

The core assumption of the CPM is that the emotion a person experiences is critically defined by the appraisal of an event. The appraisal is a component of an emotion episode that is caused by a significant event. The emotion-elicit ing event can be either internal (e.g., remembering an unpleasant interaction with a PA) or external (e.g., receiving feedback from a PA during learning with an ALT). Furthermore, the appraisal of an event is completely subjective and can therefore be inaccurate in regard to the objective situation [3].

The appraisal criteria described above can be processed on different levels. Processing can range from effortless, unconscious, low-level neural mechanisms (e.g., pattern matching for attention/perception) to complex, conscious considerations involving calculations in prefrontal cortical areas that can include propositional knowledge and cultural meaning systems [3] (for a detailed description of the levels of processing see [27]). The level of processing can also vary throughout the appraisal process, with early appraisals (e.g., relevance check) often occurring on a low uncon-
sicious level and later appraisals (e.g., normative significance) requiring conscious high levels of processing.

Even though emotion episodes are defined by the interaction of several components over time, the CPM postulates that the nature of an emotion is exclusively characterized by the pattern of appraisal outcomes and their development over time (see [3]). This implies that the CPM does not rely on a limited set of emotions. Instead this model leaves room for an infinite number of different emotion episodes without any categorical limitations (for a different perspective see [29]). However, the CPM can also predict distinct emotions based on appraisal outcomes that occur frequently. Those appraisal configurations are referred to as modal emotions [3].

Conclusions and Future Directions

This paper has afforded us the opportunity to briefly describe two of the most comprehensive models of metacognition and emotions found in the literature. We have specifically chosen to describe and succinctly present the assumptions of each one. This was purposely done to familiarize the ITS community with their comprehensiveness as well as start a discussion about the relative importance, contributions, and so forth. Due to space limitations, we raise several key conceptual, theoretical, methodological, and design issues that will serve as a major source of discussion at the ITS 2016 workshop. In this section, we focus on four specific issues: conceptual/theoretical issues, analytical issues, design of external regulating agents, and their interventions in supporting metacognitive and affective self-regulatory processes.

Some major conceptual and theoretical issues stem from the overlapping nature of assumptions, constructs, mechanisms, feedback mechanisms, and predictions in Winne and Hadwin’s IPT [1,2] model with the addition of Azevedo, Greene, and colleagues’ [19] macro-level, micro-level, and valence approach to cognitive and metacognitive processes, as well as Scherer’s [3] CPM theory. First, several issues arise when comparing these models and theories. For example, are metacognitive judgments similar to affective appraisals? Is it an issue of terminology? They both seem to rely on some evaluation of an antecedent event that triggers a response that leads to a metacognitive judgment or a multicomponential response that subsequently necessitates the use of a control strategy (e.g., making an inference) or an emotion regulation strategy (e.g., cognitive reappraisal; see [29] for a detailed explanation).

Second, while both theories provide a comprehensive set of constructs, mechanisms, and so forth, it is fruitful to consider amalgamating them in order to include a comprehensive SRL model that cohesively integrates metacognition and emotions as they are integral components required to explain complex learning. Other issues to be discussed at the workshop will include the following: (1) How does context (including type of ITS and instructional material within the ITS), as well as internal (e.g., individual differences, SRL knowledge and skills) and external (e.g., provision of adaptive scaffolding and feedback, instructional resources) conditions, impact the quality and quantity of metacognitive judgments, affective appraisals, and subsequent self-regulatory behaviors? (2) Do internal standards for monitoring cognition differ
than the ones for SECS? If so, how? Is it quality, quantity, complexity, and so forth? How are they used (for monitoring), revised, updated, and so forth prior to, during, and following learning? (3) What factors influence the sequence of processes and mechanisms specified in both models and theories? (4) Both the IPT and CPM only make assumptions about the individual. As such, how can they be extended to account for multiple agent (e.g., human–human, human–artificial agent) interactions during learning in complex contexts involving some type of computerized system(s) (e.g., high-fidelity mannequins, human–robot interactions, tangible landscapes)? (5) Do feedback mechanisms align during monitoring and regulation of cognitive and affective processes? Or are they parallel systems that “communicate” with one or several SECS or are they managed and coordinated by a complex conflict resolution mechanism? These are a few of the questions that need to be addressed by interdisciplinary researchers from the cognitive, learning, affective, social, and computational sciences.

Several analytical issues can be raised, assuming researchers move towards using multimodal multichannel data (see [6,30]). For example, can metacognitive and affective data be analyzed: (1) for distinct metacognitive and emotional signatures within channels (e.g., is the gaze pattern of frustration different from the gaze pattern of confusion? Does the gaze pattern precede, occur concurrently [or with some short latency], or follow the gaze behavior patterns?); and across data channels (e.g., is the electrodermal activity boredom pattern also observed when the boredom gaze pattern is detected? Are there short corresponding verbal or human–computer interactions that accompany the gaze behavior patterns? What does it mean if they do not match?); (2) to assess which pattern(s), both within and across channels, is/are most reliable and predictive of metacognitive and affective (and affective predicting metacognitive) processes and performance measures; (3) for indications of students’ ability to adaptively monitor and regulate their metacognitive and affective processes and other external regulating agents (e.g., PA, intelligent virtual human, peer, teacher, trainer); (4) do emotions require a longer time frame to measure than metacognitive processes (or the opposite)? Are the actual lengths of these processes the same? Typically, emotions are short-lived, so we can also have multiple emotions during one instance of a metacognitive process (within the given time span); and (5) to assess the interactions and temporal sequences among SRL processes across different contexts and phases of problem solving, conceptual understanding, comprehension, and so forth. These data can be analyzed using both traditional statistical methods (e.g., multilevel modeling) as well as data mining and machine learning techniques, including hidden Markov model analysis and hierarchical and differential sequence mining algorithms [31]. The community should adopt principles of data fusion, feature fusion, and decision fusion proposed in several leading publications.

These issues have further important implications for designing ITS with PAs or intelligent virtual humans that can detect, model, and foster students’ cognitive, affective, and metacognitive SRL processes during learning, problem solving, and so forth. When designing these external (monitoring and) regulating agents, we must address critical issues regarding timing, because the agent must be provided with the appropriate threshold in order to detect the student’s real-time monitoring (e.g., onset of a metacognitive judgment, and its associated micro-level classification and valence; for
example, “I do not understand how this diagram is related to my current learning goal” and determine how to scaffold and provide helpful feedback and individualized instruction. For example, are there multichannel data that preceded this verbalization to indicate metacognitive monitoring, based on sampling of previous behaviors (e.g., prolonged fixation on the diagram, unequal time spent reading text and less time spent inspecting the corresponding diagram, gaze behavior patterns of a specific duration indicative of Winne and Hadwin’s SMART [e.g., selecting assembling translating], and so forth? A further complication that can arise is that the verbalization is somewhat different—for example, “I know that this diagram is not related to my current learning goal.” How does verbalization (metacognitive judgment) differ from the previous example? How do these differences impact the type of external regulation needed to address this student’s specific learning need, at a particular point in time, during learning with a particular ITS? Lastly, this issue can be further complicated by a verbalization such as “I do not understand how this diagram is related to my current learning goal and I am getting very confused because I am not sure what to do next.” This utterance seems to indicate a (hopefully) correct metacognitive judgment related to the relevancy of the representation of information (i.e., diagram) provided by the ITS as well as an appraisal of the corresponding affective state with a rather nebulous statement about whether there is a lack of cognitive strategy, emotion regulation strategy, or both. In sum, besides using natural language processing (NLP) to capture these utterances in real time to make inferences about metacognition and affect, multichannel data would be necessary to provide contextual cues, history of prior SRL knowledge, behaviors, performance, learning, assessments, and so forth.

Lastly, an additional challenge when designing agents deals with the type of intervention the agent provides. The agent can intervene by prompting the student to engage in a particular cognitive, affective, and metacognitive (CAM) process or by confirming that it has correctly identified that process the student is engaging in. We consider this a challenge because to create agents that are capable of intervening by prompting we need to understand student behavior, and if we have trouble doing so as humans, due to the complexity and unpredictability of human behavior, how do we expect to program agents to be capable of doing it using algorithms that are not responsive to individual human behavior? Thus, instead of making the inferences based solely on data, the agents can be programmed to intervene by engaging in dialogues (through NLP) with the students to confirm if what they are detecting in the data is correct. In addition, instead of waiting for the right amount of data, the agent can intervene to obtain the ground truth about the student’s use of CAM processes and establish a rapport with the student to gain this ground truth. The more information the agent can obtain, the greater the likelihood it will be able to make accurate inferences regarding student behavior. Ideally, artificial agents should have access to multichannel data and be able to understand and reason from the data while determining how to adapt their own behavior to support learners’ SRL. In conclusion, we argue that conceptual, theoretical, empirical, and analytical investigation on the coupling of

---

2 This assumes that verbal data are collected and that the human is capable of producing language (e.g., coded utterance based on concurrent think-alouds) that is codable.
metacognition and emotions is necessary to provide comprehensive understanding of their roles in supporting learning, problem solving, and conceptual understanding with ALTs.

Acknowledgements. This study was supported by funding from the National Science Foundation (DRL 1431552).

References

Workshop 5

Building ITS Bridges Across Frontiers

http://its.upb.ro/index.php/organizers/
Preface for the Workshop on Building ITS Bridges Across Frontiers

Collocated with the 13th International Conference on Intelligent Tutoring Systems

Zagreb, Croatia, 2016

The goal of our workshop was to extend the community of Intelligent Tutoring Systems in two directions by build both geographical and conceptual bridges. First of all, our aim was to provide researchers with the opportunity to present their work by disregarding geographical distances and facilitating their integration in the ITS community. Consequently, the main group of participants for our workshop consists of newcomers in the ITS community from representative universities and research institutes from a diversity of countries covering the Czech Republic, Tunisia, France, Mexico and Romania.

A second target of our workshop was to bring new ideas, approaches and experiences to the ITS community. In the end, our workshop presents together four studies covering different facets of ITS with emphasis on the underlying technologies. Their authors discuss different approaches and methodologies including: a) Student models adequate for particular educational settings that are used in real world scenarios, b) Educational experiments conducted with students engaged in using an intelligent environment for learning Java that highlights the benefits of motivation play, c) A generic evaluation approach for electronic collaboration based on anomaly detection and explanation assessment that considers both technological and human components, and d) A method of generating arborescent test groups based on a genetic algorithm that best matches the degree of difficulty of underlying questions.

Ștefan Trâușan-Matu, University Politehnica of Bucharest, Romania

Stefano Cerri, University of Montpellier, France

Mihai Dascălu, University Politehnica of Bucharest, Romania
Choosing a Student Model for a Real World Application

Jiří Řihák and Radek Pelánek
Masaryk University Brno

Abstract. A student model is a key part of an adaptive educational system. Research literature offers wide choice of student modeling approaches, but little guidance on how to choose a suitable model for a real world application. Which aspects of student modeling are most important for a particular setting? How to decide whether a model is ready for application? We study these questions for a particular case study using generally applicable techniques. Our results illustrate the use of “wide and shallow” models to identify the most important aspects of student modeling. The case study (from the domain of mathematics education) specifically demonstrates that response times deserve more attention in student modeling.

1 Introduction

Intelligent tutoring systems aim at providing students with personalized, adaptively selected content. This behaviour is typically based on a student model, which estimates the current knowledge of students. The research literature provides many student modeling ideas, but little guidance on how to choose and tune a student model for particular application. For developers of real world applications it is thus difficult to choose a suitable model and this hinders spread of intelligent techniques in education.

There are many student modeling approaches and many features that we may include in our model. To illustrate the wide potential of student modeling, we mention several possible aspects of student modeling with examples of research. Basic student models focus on correctness of answers and modeling of learning in different ways (e.g., Bayesian knowledge tracing [2] or models based on logistic function [7]). Models need to know mapping between items and modeled skills and potentially relations between skills. This issue is called skill modeling [3] and can be addressed by wide variety of approaches (e.g., Q-matrix [1] or Bayesian methods). In addition to correctness of answers we can use other observational data, e.g., response times [8], common wrong answers [16], or hints usage [15]. Student population is typically not completely homogeneous and thus models may be improved by student clustering [6]. Research papers typically focus on detailed analysis of one of these aspect without any consideration of others.

From practitioner point of view there are several important questions which are not properly addressed in current literature: Which aspects of student modeling are most important for a particular setting? Which aspects are of key
importance and which provide only minor improvement not worth bothering with? Which techniques should be used to select an appropriate model? Is a student model ready for practical application? Are model parameters stable?

In this work we try to address some of these issues. As a case study we use a real adaptive educational system in its early stage of application, where the choice of student modeling approach is a real, pressing development issue. We explore a range of modeling approaches, discuss their relations and comparison, and study parameter stability. The result illustrate the use of “wide shallow” models to identify the most important aspects of student modeling for a particular application. Specifically, the results show usefulness of paying attention to response times and stability of model parameters, which are both neglected issues in current literature on student modeling.

2 Setting

To explore the issue of model selection we utilize data from a realistic case study and compare a sample of potential student models.

2.1 The MatMat System

The MatMat system (matmat.cz) is an adaptive practice application for children, which covers the area of basic arithmetic (counting, addition, multiplication, etc.), its functionality is similar to Math Garden [8]. The system is available freely online (registration is possible, but not required). The application is adaptive, its behaviour (used item selection algorithm) and default student model are described in [14].

The studied application is in many aspects typical online educational system. One specific aspect of the domain of basic arithmetic is that response times seem particularly important. For example multiplication of small numbers starts as procedural knowledge (a child knows that $3 \cdot 5$ is $5 + 5 + 5$ and is able to do the calculation using fingers) but ends as declarative knowledge (a child knows $3 \cdot 5 = 15$ without further thoughts). In both cases a child can give a correct response with high probability, to distinguish between these levels of knowledge it seems useful to utilize response times.

The currently available data from the system comprise 150,000 answers, the distribution of answers between users is highly skewed (many users answer only a few questions). The system contains examples divided into 5 high level concepts (counting, addition, subtraction, multiplication, division), each of these concepts contains around 50-700 items, over 2,000 items in total. Although the data set is relatively small (compared to many data sets currently used in research), it constitutes a realistic case study for the studied questions – developers of new systems need to make decisions on the use of student models with relatively small data.
2.2 Student Models

Our purpose in this work is not to find some “optimal” model, but rather to provide insight into importance and relation of different aspects of student models. Thus the following list of models is certainly not an exhaustive enumeration of options feasible for the particular application, but rather a selection of reasonable approaches.

**Basic Modeling Approach** As a baseline for comparisons we use a simple *item average* model, which predicts performance based on the percentage of correct answers for the particular item (ignoring student characteristics).

As the basic modeling approach, which is further studied and extended, we use student modeling based on logistic function. Note that Bayesian knowledge tracing [2], the currently dominant student modeling approach, is not suitable for the studied application – the binary skill assumption is unrealistic and incorporation of difficulty of items and timing information is possible, but complicated. The modeling approach based on logistic function naturally allows incremental increase in skill (which is expected in the domain of basic arithmetic) and can quite easily incorporate different additional aspects.

Specifically, we use the Elo rating system [4, 11], which can be seen in its basic form as a heuristic for parameter estimation of the Rasch model (it gives nearly the same parameter estimates [11]). We use the Elo rating system due to its many practical advantages (a key factor for real world application) – parameter fitting is done naturally online, it is easy to implement and extend.

The model has a student skill parameter \( \theta \) and an item difficulty parameter \( d \). The probability that a student answers correctly is estimated using a logistic function of a difference between the skill and the difficulty: 

\[
P(\text{correct} | \theta, d) = \frac{1}{1 + e^{-(\theta - d)}}
\]

After observing the student’s response the parameters are updated as follows:

\[
\theta := \theta + U \cdot (\text{correct} - P(\text{correct} | \theta, d)),
\]

\[
d := d + U \cdot (P(\text{correct} | \theta, d) - \text{correct}),
\]

where \( \text{correct} \in \{0, 1\} \) denotes whether the question was answered correctly and \( U \) specifies sensitivity of parameter estimates to the last attempt. Based on previous research [9–11] we use an uncertainty function \( U(n) = \alpha / (1 + \beta n) \), where \( n \) is the number of previous updates to the estimated parameters and \( \alpha, \beta \) are meta-parameters (fitted using a grid search: \( \alpha = 1.0, \beta = 0.1 \)).

**Domain Modeling** An important modeling aspect is domain modeling, particularly choosing the granularity of the model and the item-skill mapping. For the studied system, the items that are presented to students are for example: “5 \( \times 3 \)”, “2 + 3” (with additional visualization), or “choose a number 6 on the number line”. There are many ways how to model this domain; for the evaluation we have chosen three representatives. All of them have difficulty parameter for each item, they differ in the way they model student skills:

- **Basic model** – for each student there is a single skill.
- **Concept model** – for each student there are skill parameters for each of the 5 main concepts, these skills are assumed to be independent.
- **Hierarchical model** – skill of each student is described in tree-like structure (see Fig. 1) with 3 levels of concepts and sub-concepts (described in [14]). Due to this structure skills for different concepts are interconnected.

![Fig. 1. The structure of the hierarchical model.](image)

**Response Times** Most student modeling approaches consider only correctness of answers. Response times, however, can provide useful information about student knowledge, particularly for the studied domain. In this work we use the following approach to incorporating response times to student models. We combine correctness and response time into single performance measure \( r \). For wrong answers we ignore response times (i.e., \( r \) is constantly equal to 0). For correct answers we transform the value 1 into an interval \([0, 1]\) by one of the following functions (see Fig. 2):

- **noTime** – no use of time, \( r = 1 \).
- **thresholdTime** – response is classified as fast or slow (based on the threshold 7 seconds, the median response time in system); \( r = 1 \) for fast responses, \( r = 0 \) for slow responses.
- **expTime** – similar to the previous one; for fast responses \( r = 1 \), for slow responses \( r \) decreases exponentially (see [14] for more details).
- **linearTime** – response is linearly decreasing with increasing time (until it reaches 0); \( r = \max(0, 1 - t/14) \).

There are, of course, other possible approaches to modeling time. For example a similar approach based on the Elo rating system uses a “high speed high stakes” rule [8], which takes time into account even for wrong answers. A conceptually different approach is based on separate modeling of correctness and speed [13].
Wrong Answers  Another potentially useful source of information about student knowledge is the specific answer, i.e., in the case of incorrect answer we can look at the specific mistake that was made. Previous research for example studied differences between common wrong answers and other mistakes [16]. Analysis of answers from the studied system shows that there is large proportion of missing answers. Moreover, these answers often have very short response times and are often present in sequences (i.e., users are sometimes skipping sequences of items).

We incorporate this insight into student models in the following simple way. We compute probability of missing answer based on the number of immediately preceding missing answers of a student. Overall prediction is the product of the original model prediction and the probability of not missing an answer. This new prediction is also used in the update of model parameters.

3 Evaluation

We analyze the described models on the data from the MatMat system – we start with comparison of prediction accuracy of models and continue with more detailed analysis of parameter values.

3.1 Prediction Accuracy

For comparison of predictive accuracy of models we use rather standard evaluation setting: repeated random cross-validation (20 runs) with student stratified train/test set division (70%/30%). As the basic performance metric for comparing model we use Root Mean Square Error (RMSE), which is a standard choice in the case of predicting correctness of answers [12].

Fig. 3 gives comparison of different domain modeling approaches and of the impact of explicit treatment of missing answers. With respect to domain modeling, we see that more complex models are able to improve predictions, although increasing complexity of models brings only diminishing improvements. The figure also shows that explicit treatment of missing answers can significantly improve accuracy of some models. This very straightforward and simple
technique brings large improvement for baseline model, nontrivial impact on the basic model and the concept model, but only minimal impact on the hierarchical model, which is more flexible and can quickly adapt to the students who are just skipping items. This shows that different aspects of student modeling are partly compensatory, i.e., novel modeling ideas should not be judged (only) by their ability to improve simple baselines.

The evaluation of models which consider timing information is more difficult. Standard performance metrics [12] consider only binary information about correctness, but the point of models with timing information is to also distinguish between slow and fast responses. The full discussion of this issue is beyond the
scope of this paper, here we report just three selected metrics to provide basic insight into model behaviour. At first, we use the standard RMSE (i.e., the performance metric ignores the timing information). At second, we use “ternary” version of RMSE where observations are labeled as 0 (wrong), 0.5 (correct, slow), and 1 (correct, fast). At third, we use the Area Under the ROC Curve (AUC) metric [12], which ignores the timing information, but considers the prediction only in relative manner.

The comparison of models according to these metrics is shown in Fig. 4. With respect to domain modeling the results are similar to results presented in Fig. 3. Also note that the modeling of timing information is rather orthogonal to domain modeling. With respect to different variants of modeling response times, we see large differences between models, but these differences are not easy to interpret due to the impact of used metric. The results for the two variants of RMSE are not very surprising, since RMSE takes into account absolute values of predictions and different models are trained to predict different values, e.g., models with linearTime have much higher RMSE since they are trained to give much smaller predictions (with different meaning). The best models with respect to reported RMSE metrics are those that match the performance metric used for evaluation. The interesting part of results is the comparison with respect to AUC. Although the evaluation metric does not take response time into account, the best results are achieved using linearTime model variants. This modeling approach improves the relative order of predictions (slower students are relatively more likely to make mistake than faster students with the same overall correctness).

3.2 Impact on Estimated Parameters

The summary metrics for predictive accuracy are notoriously hard to interpret [12]. How important are small differences in RMSE values? To get insight into differences between models we analyze correlations between parameter values, particularly the item difficulty parameters, which have clear interpretation and direct impact on the adaptive system behaviour (e.g., adaptive selection of items or provided feedback for students).

Fig. 5 shows correlations between item difficulties for different models. Darker color means higher correlation, i.e., more similarity in difficulty estimates (less important difference between models). Unsurprisingly, there is a large gap between the baseline model and other more sophisticated models. The impact of domain modeling is nontrivial, but not pronounced. Different utilization of time, however, brings considerably different parameters. The degree of change is proportional to the intensity of time utilization, linearTime extension is the most different. Also note that the figure contains repetitive 4×4 pattern corresponding to different time usage. This means that domain modeling and time modeling are almost independent modeling aspects and provide change (and possible improvement) in different directions.

To provide better intuition beyond the summary evaluation metrics and correlations, we provide a specific illustration of the model impact on parameters of simple addition and subtraction items. Fig. 6 presents comparison of estimated
parameters for a model with and without timing information. The estimated parameters are only weakly correlated and there are significant differences (e.g., “1-1” and “8-5” have the same difficulty according to model without response times, but quite different difficulty according to the model with time). Particularly, note the highlighted subtraction examples of the type “X-X”, which the model with timing information systematically rates as easier than the model without time.

3.3 Parameter Stability

Before we use a model in an adaptive educational system we want to be sure that its parameters are reasonably stable. For example in Fig. 6 we can see that the estimated parameters are probably not yet completely stable (one would intuitively expect for example that the items “X-X” would either have very similar difficulty or be ordered). How to objectively judge parameter stability? How quickly do parameter values stabilize? How much do different models different in their speed of convergence? Such questions do not get much attention in student modeling. A recent exception is the proposal for multifaceted evaluation of student models [5], where authors include parameter stability as a criterion for model evaluation, but they discuss only a specific model (BKT) and do not focus on dynamics of parameter stability.

To evaluate these questions we performed the following experiment. We took two data samples $D_1, D_2$ of size $K$ (student stratified, without intersection). We consider a particular model type $M$ and train an instance $M_1$ using the data set $D_1$ and an instance $M_2$ using the data set $D_2$. Then we evaluated the correlation
Fig. 6. Comparison of estimated difficulties of selected items by the basic model with and without time usage. Note that absolute values of difficulties are not entirely comparable because models are trained to predict different values.

of parameters of $M_1$ and $M_2$ (we use only item difficulty parameters since the data sets are student stratified).

Fig. 7 shows increase in parameter stability with the number of answers used for training the model. The figure compares different domain modeling approaches and different time uses. The left part of the figure shows that the hierarchical model is slightly more stable than simpler models as it can better carry information across items. The right part of the figure shows high increase in stability of models which utilize response times. Fig. 6 provides a specific illustration, note that the group of similar items of the form “X-X” have very similar difficulty according to the model utilizing response time, but widely different difficulties for model without response times. This increase in stability (resp. faster convergence) is probably mainly due to the use of more “bits of information” per each answer. Increase in stability is also proportional to intensity of time utilization – linearTime extension makes the most significant use of response times.

4 Discussion

For the studied case study, the main conclusion is that differences in modeling of response times have larger impact than differences in domain modeling. The results give actionable insights for practical application and directions of further analysis (e.g., separate treatment of correctness and speed may be useful).

From the wider perspective, the case study highlights issues with practical applications of research results. The focus of current research is mostly on details of one particular aspect of student modeling (e.g., domain modeling or different ways of modeling learning). For applications of student models in real
world contexts, we need more focus on relations of different aspects of student modeling and on their combinations. Our results show that incorporation of different aspects of student modeling (even in simple way) may be more important than detailed modeling of one particular aspect. Different modeling aspects are not, however, always orthogonal. An approach that brings large improvement over baseline (when studied in isolation) may have negligible impact when incorporated into a more complex model – in our experiments this is the case of treatment of skipped answers. From application point of view, more attention also needs to focus on “production readiness” of models.

These concerns are not just pragmatic, development issues, they have important consequences for research, e.g., comparison with baseline not sufficient to show that a modeling technique will improve also more complex models. They can also inspire interesting scientific questions. Can we devise compositional modeling approaches allowing simple combination (and evaluation) of different aspects of student modeling? To what degree is the importance of individual aspects domain specific? Can we formulate any generally applicable guidelines? What is the best way to study stability of model parameters and how can it help in model comparison?

References


5. Yun Huang, José P González-Brenes, Rohit Kumar, and Peter Brusilovsky. A framework for multifaceted evaluation of student models. In Educational Data Mining, 2015.


Experiments with Students using an Intelligent Environment for Java

1María Lucía Barrón Estrada, 1Ramón Zatarain Cabada, 2Raúl Oramas Bustillos, 3Giner Alor-Hernández,

1Instituto Tecnológico de Culiacán, Culiacán, Sinaloa, México, 2Universidad de Occidente, Campus Culiacán, 3Division of Research and Postgraduate Studies, Instituto Tecnológico de Orizaba.

{lbarron, rzatarain}@itculiacan.edu.mx, raul.oramas@udo.mx, galor@itorizaba.edu.mx

Abstract. This paper presents the results of high-school students using a learning environment, named Java Sensei, for learning Java. We first give a brief introduction of the web learning environment for studying Java and a description of the experiments we made with the tool. Later, we show the experiments made with different students studying a first programming language. In our analysis of results, we found that the ILE did not have an impact on high-school students. We consider that motivation play an important role when learning a first programming language.

Keywords: Intelligent Learning Environments, Intelligent Tutoring Systems, Affective Computing, Programming Languages.

1 Introduction

Learning computer programming is a challenge for many students because they need to make an effort to understand new concepts like statements, expressions, control structures, data structures, etc. in order to create programs that solve concrete problems [1]–[3].

Other factors that contribute to this challenge are the teaching methods because teaching is not personalized and teacher’s strategies usually do not support all student’s learning styles; moreover, teachers are more concentrated on teaching a programming language and its syntactic details. On another hand, students adopt incorrect study methodologies because programming should be essentially practical and very intensive, quite different from what is required in many other courses (more based on theoretical knowledge, implying reading and some memorization). Programming demands intense extra classes work [4].

Previously, we introduced an Intelligent Learning Environment (ILE) called Java Sensei [5] which performs detection of affect analyzing several features from student's faces.
The idea behind the ILE Java Sensei is to have a flexible and interactive Web-based learning environment that considers the cognitive and affective states of students in order to improve the skills of design, creation, and execution of programs written in the Java programming language. In this paper, we present a descriptive statistical analysis study about the impact of using the ILE Java Sensei to develop programming skills using Java.

The paper is organized as follows: section 2 shows related works, section 3 describes the Java Sensei learning environment. The experiment description is shown in section 4, and its results are analyzed in Section 5. Conclusions and future work are discussed in Section 6.

2 Related works

Several ITSs and ILEs have been implemented in the programming domain to develop the student’s programming skills. Among these we found PROUST[6], LISP Tutor[7], JITS[8], BITS[9], J-LATTE[10], iList[11], OOPS[12], JavaGuide[13], Nooblab[14], PHP ITS[15] and the JavaTutor System[16] which were developed to teach different programming languages. PROUST and the LISP Tutor were presented in the 80’s and can be considered as two early ITSs in the programming domain.

Different IA techniques and approaches were used in each work. For example, J-LATTE uses a Constraint-Based Modelling (CBM) for representing the domain knowledge as a set of state constraints. A constraint specifies certain conditions that must be satisfied by all correct solutions. JavaGuide is a system which guides students to select appropriate questions analyzing the effect of personalized guidance while learning to program in Java. The JavaTutor System, is a multimodal affective tutor that uses dialogs with machine learning techniques. The JavaTutor uses different specialized sensors for emotion recognition and for adapting the learning activities of the student.

We made a comparative analysis for these ITS and ILEs, and one of many advantages of Java Sensei over these similar systems is that this system is focused on the improvement of good programming practices. For example, when the student finds a solution to a problem, the system recommends a better or optimal solution. Another advantage of Java Sensei is a recommendation system that uses the student model to customize exercises and assistance depending on the type of user (student). Regarding the use of technologies, the main advantage of Java Sensei is that can be used on any kind of computer (PC, laptop, mobile) that supports a browser and have installed a camera and a microphone, which is common today. Unlike other similar systems, Java Sensei needs no special sensors for emotion recognition.

3 Description of Java Sensei Environment

Java Sensei is an ILE for learning Java programming. The ILE contains an Affective Tutoring System (ATS) working in a Web environment (http://javasensei.ddns.net/). The Java Sensei Environment contains seven layers (see
Fig. 1: Web Layer, Service Layer, Server Layer, Domain Layer, Student Layer, Data Management Layer, and Data Layer that works together to create a Web-based environment to teach the Java programming language.

![Java Sensei Architecture](image)

Fig. 1, Java Sensei Architecture.

To get full access to the ILE, the student must have a Facebook account and a PC, laptop, or mobile device with a browser, allowing the use of the Webcam. Fig. 2 shows the Web environment that represents the student’s interface with a menu organized by lessons; the icons on the top right corner let the student know their progress and see some resources about Java.

The system uses a set of files in JSON format for knowledge building. Knowledge representation is build using the Knowledge Space Theory (KST)[17]. The domain model represents basic skills that students must master and is modeled by a graph representing the knowledge space. These skills are organized into lessons: Introduction to Java, Variables and Calculations, Selection, Iteration, Methods and Arrays.
Each lesson presents 15 exercises to the student, who must solve them in order to learn and practice the topics of Java language. Fig. 3 shows a representation of knowledge building. The nodes in level one represent the lessons in the domain model and their child nodes denote one or more exercises that the student must solve.

Each exercise was implemented using the pedagogical model known as "problem-solving" [18] where the student learns as he solves problems with a certain structure. The system uses three different strategies for solving problems. Strategy 1 is used to evaluate theoretical concepts allowing the student to answer “true-false” exercises. Strategy 2 presents a complete program or piece of code and ask the student to determine the output of the code. Strategy 3 is a combination of the two previous strategies and is used to create more complex exercises involving a number of steps to reach the solution.

Fig. 4 shows how an exercise was built with the ITS Engine module that implements the solving problem strategy using an Example-Tracing Technique[19][20]. Example-Tracing has the advantage of providing step-by-step guiding to the students and providing with multiple strategies to the problem solution, including an optimal
or sub-optimal solution or misconception management. Additionally, the ITS Engine module makes an integration of cognitive and affective information in a behavior
graph which is used by a pedagogical agent in order to show affect and empathy to the student while using the ILE.

1. The system shows the exercise to the student
2. The student selects an Incorrect option
3. The student selects the correct Option. Ends of the exercise.

Fig. 4. Exercise with example-tracing technique.

The student model represents cognitive and affective information of the student and it is recording each student’s action in the exercise while the student uses the tutor. For example, when the student solves an exercise the system works with an affective module that uses a fuzzy logic system and a back-propagation neural network for emotion recognition.

The emotion recognition system was built in three steps: the first one extracts features from faces images building a corpus used to train the neural network. The second step was the implementation of the neural network. We used the Java-based algorithms implemented in NeuroPH[21] to classify the emotions by using a neural backpropagation network. The third step integrate extraction and recognition into a fuzzy system. For training and testing the neural network, we used the corpus RAFD (Radboud Faces Database), which is a database with 8040 different facial expressions that contains a set of 67 models including men and women.

The student’s emotion is read by a Pedagogical Agent (see Fig. 5) that is responsible for transmitting empathic messages as a human tutor would. The agent represents the part that humans perceive as an emotion. The pedagogical agent has been provided with both facial expressions and dialogues in order to interact with the student. We use the Ekman theory[22] to recognize emotions of the students.
4 Description of the experiment

We carried an experiment in December 2015 with 32 students (15-16 years old) from High School (Instituto Chapultepec) from Culiacán, Sinaloa in México. The group was learning JavaScript as first programming language and they do not have background programming with Java.

The experiment was divided in three 45-minute sessions. In the first session, the students solved a pre-test on paper. In the second session, the students practiced with the ILE Java Sensei, and in the last session, the students solved a post-test and answer a survey about the ILE.

In this experiment we wanted to find out if the student’s knowledge improves using the ILE Java Sensei while learning four topics: Introduction to Java, Variables and Calculations, Selection and Iteration Control Flow because they are the main topics in any introductory programming course.

Pre and post-test questions were designed based on the content of the ILE Java Sensei and contained 15 questions each.

5 Results of the experiment

In this section, we present the results of the experiment and the survey.

5.1 Survey results

The survey has eight multiple option questions. The students were asked to express their opinion for each question using the 1 - 5 Likert scale, where 1 means strongly disagree, 2 disagree, 3 undecided, 4 agree, and 5 strongly agree. The survey was
anonymous and the same number of students participated using the ILE and answering the survey (n=32).

The questions from the survey and responses are presented in Table 1. According to the obtained survey results, for question one, we found that about 60% think the ILE Java Sensei seems easy to use, 6 students (18.7%) strongly agreed with the facility of use of the ILE’s user interface and 14 students (43.7%) also agreed. For question two, we found that 16 students (50%) were neutral about the experience with the ILE but 37% enjoyed using the ILE and only 12% dislike using the environment. In question 3, 13 students (40%) found easy to learn with the ILE and other 40% was neutral. In question 4, 16 students (57%) found attractive to work with the ILE more than a traditional session in the classroom and 9 of these students (20%) were neutral. For question 5, 17 students (53%) found useful the information of the ILE, 30% of the students found confusing some of the questions/answers in the ILE as can be seen in question 6, but almost 60% did not have problem to understand the questions. Most of the students think the system time response is fast (question 7) and almost 50% would like to use the system again (question 8). These results lead us to think that students were satisfied with the ILE Java Sensei.

Table 1. Questions from the survey and given responses

<table>
<thead>
<tr>
<th>Num.</th>
<th>Question</th>
<th>Number of responses for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The ILE seems easy to use</td>
<td>3 1 8 14 6</td>
</tr>
<tr>
<td>2</td>
<td>I enjoy using the ILE</td>
<td>0 4 16 8 4</td>
</tr>
<tr>
<td>3</td>
<td>I found it easy to learn Java using the ILE</td>
<td>0 6 13 11 2</td>
</tr>
<tr>
<td>4</td>
<td>Using the ILE is more constructive than taking a traditional class</td>
<td>2 5 9 12 4</td>
</tr>
<tr>
<td>5</td>
<td>The ILE provides me the data I need to remember</td>
<td>1 6 8 13 4</td>
</tr>
<tr>
<td>6</td>
<td>The answers of the system are clear</td>
<td>4 4 5 15 4</td>
</tr>
<tr>
<td>7</td>
<td>The response time of the system was fast</td>
<td>1 5 7 10 9</td>
</tr>
<tr>
<td>8</td>
<td>I liked the system and I would use it again</td>
<td>1 3 13 7 8</td>
</tr>
</tbody>
</table>

5.2 Experimental evaluation

In this experiment we wanted to find out if using the ILE Java Sensei leads to improvements in students’ knowledge/skills (i.e. scores). Then, our first condition was the student’s performance scores in the pre-test exam without having studied the ILE Java Sensei. The second condition was the student’s performance scores in post-test exam having studied the ILE Java Sensei. Each student participates in both conditions of the experiment. Fig. 6 shows differences between the results that we got in a pre-test and post-test evaluation.
**Fig. 6.** Pre-test and post-test results.

For these datasets, we applied a statistical test called: Paired Samples T-Test because each student participates in both conditions of the experiment. Table 2 shows the first part of the results.

<table>
<thead>
<tr>
<th>Table 2. Paired Samples Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PAIR 1</strong></td>
</tr>
<tr>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td><strong>N</strong></td>
</tr>
<tr>
<td><strong>Std Deviation</strong></td>
</tr>
<tr>
<td><strong>Std Error Mean</strong></td>
</tr>
</tbody>
</table>

In Table 2 we can observe we don't have any missing values on the test variables since N=32, and also the Post-Test mean (M=37.34, SD, 10.90) is lower than the one in Pre-Test (M=47.28, 12.27). Table 3 shows the second part of the results.

<table>
<thead>
<tr>
<th>Table 3. Paired Samples T-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Paired Differences: pre-test &amp; post-test</strong></td>
</tr>
<tr>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td><strong>Std Deviation</strong></td>
</tr>
<tr>
<td><strong>Std Error Mean</strong></td>
</tr>
<tr>
<td><strong>Lower</strong></td>
</tr>
<tr>
<td><strong>Upper</strong></td>
</tr>
<tr>
<td><strong>T</strong></td>
</tr>
<tr>
<td><strong>df</strong></td>
</tr>
<tr>
<td><strong>Sig (2-tailed)</strong></td>
</tr>
</tbody>
</table>

* 95% CI for Mean Difference

Table 3 shows the results that we got when we compared our two conditions of the experiment. Mean equals to 9.93 indicate that there was a difference in the scores obtained for Pre-Test (M=47.28, SD = 10.20) and Post-Test (M=37.34, SD = 10.90) conditions. These results, t(31)=4.49, p = 0.00, suggest that the ILE Java Sensei does not have an impact on improving the students’ knowledge/skills (i.e. scores).
6 Conclusions and future work

The results of this experiment showed us that Java Sensei does not affect the performance of the high school students learning Java with null or little experience, but we would like to make more experiments controlling all the variables because in this experiment the ILE was used without some of its features. Pictures of the students were not captured because the computers used did not have Webcam, making impossible to use and test the Pedagogical Agent with emotions. We think that the experiment was developed in an inappropriate time (near to the end of the semester) because the students were focus on final exams. However, these observations need further experimentations for confirmation or refutation. In the other hand, that survey results have helped us in deciding how to focus on the future work and how to continue the development of the ILE Java Sensei.

Furthermore, in this experiment, we learned about the process of designing experiments and how to make a plan for it. Right now, we are planning new improvements in our investigation because we want to start new experiments to explore and measure more concepts like the emotional behavior of the student.

7 Acknowledgment

The work described in this paper is fully supported by a grant from the TECN, and CONACYT in México.

8 References


Towards the evaluation of electronic collaboration

Raoudha CHEBIL¹, Wided LEJOUAD CHAARR¹ and Stefano A. CERRI²

¹ The Tunisian Laboratory of Computing, web services composition, Optimization, Supervision, simulation, Multiagent systems and mobility for Outstanding Systems (COSMOS) National School of Computer Studies (ENSI) – Manouba University Campus of la Manouba, 2010 Manouba, Tunisia {raoudha.chebil,wided.chaari}@ensi-imaa.tn

² The Montpellier Laboratory of Informatics, Robotics, and Microelectronics (LIRMM) University of Montpellier 2 & the National Center for Scientific Research 161, Rue Ada - F-34095 Montpellier Cedex 5, France stefano.cerri@lirmm.fr

Abstract. The use of technologies in collaboration processes entitled e-collaboration, has solved some of its problems, but it introduced several others bringing a continuous need to evaluate them. A review of the existing e-collaboration evaluation works showed two principle disadvantages. The first consists in an unjustified neglect of human e-collaboration component in favor of its technological one. The second consists in a limited applicability of several evaluation works due to their dependence on the considered domain. Face to these weaknesses, our purpose is to propose a generic e-collaboration evaluation approach that considers both technological and human e-collaboration components. Given the particularity of the studied environments, our contribution started by a preliminary work to identify a generic e-collaboration success criterion focused on “Result Adequacy” and used in the definition of a generic e-collaboration evaluation approach based on anomaly detection and explanation. The proposed approach is applied on a collaborative e-learning scenario.

Keywords: E-collaboration; E-learning; Evaluation; Human reliability error analysis method, Anomaly detection and explanation.

1 Introduction

The technology involvement in collaboration has produced the electronic collaboration concept called e-collaboration and defined by: “the collaboration between persons engaged in a common task using electronic technologies” [9]. Actually, e-collaboration is increasingly invading our everyday lives through its multiple applications such as: e-learning, serious games, virtual meetings and e-business. In order to encourage the use of this concept and to improve its results, several evaluation works have been proposed.
Before studying the existing e-collaboration evaluation works, let us remember some constraints which consideration favors the evaluation reliability. In fact, an e-collaboration environment involves human participants having various psychological and social characteristics as well as computers having different technological properties. So the evaluation of an e-collaboration scenario must consider its human as well as its technological components. In addition, and given the large number of e-collaboration applications, it is hard to propose a dedicated evaluation for each one; so generic evaluations are interesting in this context. These findings motivated the following presented work consisting in the proposition of a generic e-collaboration evaluation work focused on both human and technological features.

This paper will be organized as follows. In section 2, we give an overview on the existing e-collaboration evaluation works and we present the current work objective. In section 3, we describe the suggested two level-based evaluation approach. To be more concrete, we present, in section 4, an application of the proposed approach on a collaborative learning scenario.

2 Literature review

E-collaboration works vary according to different properties such as the evaluated aspects and the application conditions [2]. Thanks to a bibliographic study, we were able to distinguish two groups of e-collaboration evaluation works. The first group is composed by a large number of generic evaluation works (independent from the e-collaboration application domain) focused on the e-collaboration technological component, frequently named groupware [1],[3],[10]. These works are generally based on: the inspection of several groupware criteria [3], the feasibility of a set of tasks [10] or the prediction of representative scenarios [1]. All of them present a limited consideration of the human impact on the e-collaboration. For example the work referenced by [3] proposes a groupware evaluation method inspired from Human-Computer Interaction works. It suggests a set of evaluation heuristics adapted to collaboration processes, allowing the detection of groupware problems by checking if it verifies a number of fixed criteria. In order to improve the considered groupware, the detected problems are classified and solutions are proposed.

Contrary to the first group, the second one is not generic; it focuses on evaluating specific e-collaboration applications such as e-learning [13][15] and virtual meetings [11]. Some works of this group [11] consider both technological and human criteria in the evaluation. For example, the work referred by [11] is focused on developing an evaluation support for meeting tools. For this purpose, the authors proposed a number of success criteria related to the meeting tool as well as the human participants.

Given the previously described e-collaboration characteristics and the consequent constraints introduced in the evaluation, we believe that each one of the two distinguished work groups presents a weakness. The first work group has the advantage of being generic, but it has the weakness of focusing the evaluation on general criteria mostly related to the e-collaboration technological component ignoring the human one despite its impact on the e-collaboration progress and results. In the second group,
few works have the advantage of considering both technological and human aspects in the evaluation but all of them present an important disadvantage consisting in a limited applicability to particular e-collaboration scenarios [13][15]. Consequently, our objective, in this work, is to gather the advantages of the two described work groups by proposing a generic approach which considers both technological and human e-collaboration components in evaluation.

3  
**Generic evaluation approach based on anomaly detection and explanation**

Thanks to a previously carried out work consisting of an e-collaboration modeling and simulation [4][5] and given the announced aim, we were able to identify an e-collaboration success criterion named “the result adequacy” expressing the matching degree between the obtained result and the expected one. This criterion is generic as it is independent from the considered application domain: it is exploited to propose a generic evaluation approach based on two levels consisting of an anomaly detection and explanation.

3.1  
**First level: anomaly detection**

Since the evaluated aspect consists in the result adequacy, this level starts by comparing the expected result to the obtained one by verifying the accomplishment of the fixed subgoals as well as the global goal. According to a previously carried out modeling [4], we note the important impact of e-collaborators presence as well as the technical support choice on goals completion. So we compare also, in this level, their expected and obtained states. Consequently, the first evaluation level can be summarized in the comparison between the expected and obtained states of the four principal elements constituting any e-collaboration scenario and described by the following definition:

\[ S_c = \{ E, S_0, S_{in}, G_G \} \]

In this definition, \( S_c \) denotes the e-collaboration scenario, \( E \) refers to all the e-collaboration members entitled e-collaborators, \( S_0 \) is the technological support of all the implemented interactions during the e-collaboration session, \( S_{in} \) represents the set of e-collaboration sub-goals and \( G_G \) indicates the global goal. In what follows, the term “e-collaboration context” denotes the state of the four e-collaboration elements.

In the following, the term anomaly represents a mismatch between the expected and the obtained e-collaboration elements. In order to capture the scenario progress details and especially to detect anomalies, the evaluator is supposed to recover information through the first part of the questionnaire available in the URL at the bottom of this page. This questionnaire is divided in two parts: its first part is used in this evaluation level and its second part is exploited in the following level. The recovered information, in this level, concerns the expected and the obtained states of the follow-

---

1. [https://docs.google.com/a/emi-arep.a/form/d/1HoXYW3LIVdUtbJ1_BGhFHTb0yk7p8_t6-Fbl_2k/viewform](https://docs.google.com/a/emi-arep.a/form/d/1HoXYW3LIVdUtbJ1_BGhFHTb0yk7p8_t6-Fbl_2k/viewform)
ing elements: e-collaboration global goal, the e-collaboration platform, the list of e-collaborators supposed to attend the considered e-collaboration session, and the detailed list of fixed sub-goals as well as the list of e-collaborators responsible of each of them.

To ensure the adequacy and completeness of the collected information, the questionnaire should be sent to multiple e-collaborators having different roles and different visions of the required work. The collected answers to the sent questionnaire allow the evaluator to detect the gap between the expected and the obtained contexts giving the result anomalies having to be explained in the second evaluation level. An application example of this evaluation level is presented in section 4.1.

3.2 Second level: anomaly explanation

As explained at the beginning of this paper, human behavior has an important impact on the e-collaboration progress and results. Starting from this finding, we adopted the following principle in the explanation of the occurred anomalies: each observed anomaly can be considered as a human failure due to inappropriate human or technological causes detected from questionnaire responses. According to this thinking, we use, in the second evaluation level, a human reliability error analysis method entitled CREAM (Cognitive Reliability and Error Analysis Method) [8] and adapt it to the e-collaboration environment. The CREAM method permits to deduce the possible causes explaining the occurrence of each detected anomaly as well as the causal links connecting them. This method is composed by three steps:

1. The first step consists in describing the e-collaboration progress conditions using the Common Performance Conditions called CPC [8]. As shown in Table 1, the CPC is a set of items fixed by the CREAM method and supposed to capture the principle aspects impacting the collaborative work.
2. The second step consists in identifying the anomalies called phenotypes in the CREAM method.
3. The third step is an iterative step consisting in determining the possible causes of a phenotype occurrence as well as their causal links basing on states of the previously described CPC. At each iteration, a cause and an effect are identified; in the next iteration, the last cause becomes an effect.

To avoid confusions, the term “phenotype” is usually used in the beginning of the iterative step. The terms “effect” and “cause” are usually used at the end of this step. While the intermediary effects appearing during the third step are named consequences and the intermediary causes are entitled antecedents.

Hollnagel [8] divides the antecedents into three categories according to their relation to persons (category P), to technology (category T) or to the organization (category O).

As we can notice, the second step which consists in detecting the phenotypes is carried out in the first level of our approach. In addition, some of the CPC suggested by the CREAM method are already recovered in the first evaluation level. According to these findings, the application of CREAM method was adapted to the studied con-
text by escaping the second step of the original CREAM method and by adjusting the CPC having to be recovered in this evaluation level. The first part of the questionnaire used in the first evaluation level was extended with a second part available in the same URL constituted by other issues intended to recover the CPC related to the second evaluation level.

Table 1. The Common Performance Conditions proposed by the basic CREAM method

<table>
<thead>
<tr>
<th>CPC name</th>
<th>Level descriptors</th>
<th>Expected effect on performance reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequacy of organisation</td>
<td>Very Efficient</td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Efficient</td>
<td>Not Significant</td>
</tr>
<tr>
<td></td>
<td>Inefficient</td>
<td>Reduced</td>
</tr>
<tr>
<td></td>
<td>Deficient</td>
<td>Reduced</td>
</tr>
<tr>
<td>Working conditions</td>
<td>Advantageous</td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Compatible</td>
<td>Not Significant</td>
</tr>
<tr>
<td></td>
<td>Incompatible</td>
<td>Reduced</td>
</tr>
<tr>
<td>Adequacy of MMI and operational support</td>
<td>Supportive</td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Adequate</td>
<td>Not Significant</td>
</tr>
<tr>
<td></td>
<td>Tolerable</td>
<td>Not Significant</td>
</tr>
<tr>
<td></td>
<td>Inappropriate</td>
<td>Reduced</td>
</tr>
<tr>
<td>Availability of procedures/ plans</td>
<td>Appropriate</td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Acceptable</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Inappropriate</td>
<td>Reduced</td>
</tr>
<tr>
<td>Number of simultaneous goals</td>
<td>Fewer than capacity</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Matching current capacity</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>More than capacity</td>
<td>Reduced</td>
</tr>
<tr>
<td>Available time</td>
<td>Adequate</td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Temporarily inadequate</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Continuously inadequate</td>
<td>Reduced</td>
</tr>
<tr>
<td>Time of day</td>
<td>Day-time (adjusted)</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Night-time (unadjusted)</td>
<td>Reduced</td>
</tr>
<tr>
<td>Adequacy of training and expertise</td>
<td>Adequate, high experience</td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Adequate, limited experience</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Inadequate</td>
<td>Reduced</td>
</tr>
<tr>
<td>Crew collaboration quality</td>
<td>Very efficient</td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Efficient</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Inefficient</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Deficient</td>
<td>Reduced</td>
</tr>
</tbody>
</table>

The adjustment of this method to e-collaboration context gives several inferences for each detected anomaly. Each inference is constituted by a set of explanations related by a causal-effect relationship. For better visibility of the causal link, we pro-
pose to represent the different explanations of each anomaly by a causal graph as shown in Fig. 1.

![Causal graph example](image)

In the previous example, the phenotype P1 has three antecedents A, B and C belonging to different categories. These antecedents themselves are considered as consequences in the next iteration and have different antecedents D, E and F. The arrows represent the causal links and the leaves represent the phenotype causes. In this simplified example, the phenotype occurrence can be explained by the following inferences:

\[(G \rightarrow E \rightarrow A \rightarrow P1), (G \rightarrow E \rightarrow C \rightarrow P1), (H \rightarrow D \rightarrow A \rightarrow P1) \text{ et } (H \rightarrow F \rightarrow B \rightarrow P1).\]

As shown in Fig. 1, the causal graph gives different explanations of a phenotype; but it does not show the most plausible of them. Consequently, we were interested in the application of a probabilistic approach adapted to the graph characteristics as well as the antecedent classification into categories (P, T, O). In this context, a classical probabilistic approach application attributes a probability to each explanation regardless to its category and its antecedents. Contrariwise, the evidence theory, introduced by Dempster [6] and enhanced by Shafer [12], has the great advantage of allowing to assign a belief mass to a whole set of related assumptions. Since an inference of a phenotype occurrence represents a set of explanations; this theory was judged the most adapted to our context and was applied to compute the explanations beliefs (called also plausibility).

**Evidence theory application.**

The evidence theory [6] [12] is based on the idea of attributing a belief mass to a set of assumptions. Let us consider the assumptions set \( \Omega \) called “frame of discern-
ment". The set entitled $2^\Omega$ includes the $2^n$ possible disjunctions of $\Omega$, in addition to the singleton assumptions as shown in equation (1).

$$2^\Omega = \{ A \in \Omega \rightsquigarrow \{ \emptyset, \{H1\}, \ldots, \{Hn\}, \{H1, H2\}, \ldots, \Omega \}$$  \hspace{1cm} (1)

The evidence theory distributes the belief mass (also called mass) on the elements of the $2^\Omega$ set using the $m_{\Omega}$ function defined by the following equations:

$$m_{\Omega} : 2^\Omega \rightarrow [0, 1]$$  \hspace{1cm} (2)

$$m_{\Omega}(\emptyset) = 0$$  \hspace{1cm} (3)

$$\sum_{A \subseteq \Omega} m_{\Omega}(A) = 1$$  \hspace{1cm} (4)

The previously defined function $m_{\Omega}$ differs from a probability function by the fact that the whole mass is distributed not only on the singletons, but on the combined assumptions. The adequacy of this principle to the context in question, justifies the choice of the evidence theory for computing the plausibility of the different inferences explaining a phenotype.

To be able to apply the evidence theory on the causal graph, each of the three categories must have a weight reflecting its influence on the phenotype occurrence. In the current context, there is no preliminary hypothesis expressing different impacts of the antecedent categories on the phenotype occurrence. So we assigned the same weight ($1/3$) to all the categories.

In the studied context, the frame of discernment corresponds to the set of graph nodes representing different explanations of a phenotype. According to the previous presentation of the evidence theory, the set $2^\Omega$ is composed by all the possible disjunctions of the different phenotype explanations. But in the current context, we consider only the significant disjunctions presenting a causal link between each other and constituting a possible inference of the occurred phenotype.

To compute the belief masses of the graph nodes, we applied the principle used in the work referenced by [7], defining the mass propagation in a causal graph. In this work, the mass of an antecedent depends on its consequences masses as shown in equation (5). According to this equation [7], each consequence of an antecedent 'a' gives it a part of its mass modulated by the weight-category of 'a'. The addition of these different consequences parts gives the antecedent mass.

$$m(a) = p(C(a)) \times \sum_{b \in Cons(a)} \frac{m(b)}{\sum_{v \in P, P, p(i) \times nb}}$$  \hspace{1cm} (5)

Where $m(a)$ is the mass of the antecedent $a$, $C(a)$ is the category of $a$, $Cons(a)$ is the set of consequences of $a$, $p(i)$ is the weight of category $i$, $nb$ is the number of antecedents of $b$ belonging to the category $i$.

The mass computing process starts by affecting a mass $m$ to the phenotype ($m_i$). Then an iterative mass propagation to the antecedents using the equation (5) is carried out. At the end of the calculation step, we can verify that the sum of leaves belief masses is equal to 1. This propagation is fully applicable since the causal graph is acyclic.
In this section, we presented a two-level based evaluation approach. The first level ensures an intuitive assessment of e-collaboration progress and is limited to anomaly detection. The second focuses on explaining the detected anomalies by combining the CREAM method, the causal graph concept and the evidence theory. To make the previous description more clear, we present in what follows, an application of the proposed approach on a collaborative e-learning scenario based on a social network.

4 Approach application

In order to apply and test the proposed approach and given the actual dynamicity of political scenes in many countries, we were interested in a collaborative learning scenario executed in order to enhance the political culture in countries of the “Arabic Spring”. In fact, due to a long dictatorship, it was noted that persons in revolution countries have a poor political culture. Even persons with a high intellectual level are unable to hold an adequate political discussion to interpret events and to defend their political choices. Consequently, a voluntary politician created a collaborative learning group supported by the social network Facebook to instruct good practices in the political environment. The group creator started by inviting some of his friends as well as family members supposed to expand the group by inviting other participants. He planned to have a dynamic group based on discussions, sharing news and launching events. The group creator’s role consists in explaining and/or correcting incoherent discussions and interpretations. The group presentation was written on the group page as follows: “the objective of this group consists in learning how to follow, to interpret and to react to political events. Thank you for sharing interesting and reliable information. Thank you for inviting persons who may benefit from our group and have interesting and efficient feedbacks which argue in favor of the country.”

After a period of two weeks, the observed group evolution was very weak according to what was expected. To explain this gap between the expected and the obtained contexts, we applied the previously presented evaluation approach. In order to detect and explain the e-collaboration anomalies, we started by diffusing the previously presented evaluation questionnaire. The analysis of the obtained responses was facilitated by the graphical summary available in the URL shown at the bottom of this page.

4.1 First level

Since this level is only responsible of e-collaboration anomalies detection, we focus here on the first part of the questionnaire in the purpose to realize a comparison between the expected and the obtained e-collaboration contexts as shown in Table 2.

The carried out comparison summarized by Table 2, shows that all the detected anomalies can be reduced to only one problem (or phenotype) consisting in an insuf...
icient and weak e-collaborators activity in the considered group. The second evaluation level is responsible of determining the possible causes of this problem.

**Table 2.** Comparison between the expected context and the effective context

<table>
<thead>
<tr>
<th>E</th>
<th>Expected context</th>
<th>Effective context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>At least 50 members</td>
<td>20 members</td>
</tr>
<tr>
<td>S_8</td>
<td>Facebook</td>
<td>Facebook</td>
</tr>
<tr>
<td>S_9</td>
<td>1- Invite persons to extend the group.</td>
<td>1- The group was not extended.</td>
</tr>
<tr>
<td></td>
<td>2- Share recent and interesting political news.</td>
<td>2- Too few information was shared.</td>
</tr>
<tr>
<td></td>
<td>3- Create events to react to political news and decisions.</td>
<td>3- No events have been created.</td>
</tr>
<tr>
<td></td>
<td>4- Launch discussions on current topics.</td>
<td>4- Only one discussion was launched.</td>
</tr>
<tr>
<td></td>
<td>5- Supervise the group activity, correct it and give advices.</td>
<td></td>
</tr>
<tr>
<td>G_8</td>
<td>Learn to people having a correct political behavior.</td>
<td>The goal was not attained</td>
</tr>
</tbody>
</table>

**4.2 Second level**

The analysis of the obtained responses available in the previously shown URL allowed us to detect the following causes behind the weak group activity:

4. An ambiguous understanding of the group objectives as well as the expected results, identified from large differences between responses to the question 1. From the obtained responses, we can also notice that there are no responses referring to the learning objective.

5. A low satisfaction with the working conditions detected from responses to questions 6 and 1.  
6. An availability problem detected from questions 8 and 9.  
7. An access problem to Facebook functionalities detected from responses to question 2.

These problems can themselves have multiple antecedents: it is what justifies the iterative progress of the evaluation work in this level.

8. The first problem can be caused by an inadequate group supervision consisting of: an insufficient explanation of the group interest as well as its objectives and a poor presentation of the activities to perform in the group as well as the expected results (can be deduced especially from responses to questions 4 and 7).

9. The second problem can be due to an inadequate group supervision (described below) if the group objective and the expected results were not correctly highlighted.
It can also be caused by members disinterest in political topics (can be deduced especially from responses to questions 4, 7 and 10).

10. The third problem of availability can be due to a period inadequacy which prevents e-collaborators to spend time on doing the required actions and consisting in a work overload period or on the contrary, a holiday period (can be deduced especially from responses to questions 8 and 9).

11. The fourth problem of access to the available functionalities can be caused by an interface inadequacy, limited internet connectivity or limited technical skills (can be deduced especially from responses to questions 2 and 6).

For more clarity, the causal link between the different antecedents is illustrated by the causal graph in Fig. 2. According to the Hollnagel classification scheme, the antecedents we have just distinguished belong to the three different categories: P, O and T. The belief masses of the different phenotype explanations are calculated according to the evidence theory expressed in equation (5).

![Causal graph explaining the occurrence of the phenotype “weak activity”](image)

Fig. 2. Causal graph explaining the occurrence of the phenotype “weak activity”

To clarify the application of the evidence theory, especially equation (5), we explain, in the following, the calculation of the belief mass of the cause “Inadequate group supervision” referred by ‘a’. This cause has two consequences “Understanding problem” referred by ‘b1’ and “Low satisfaction” referred by ‘b2’. In this case, the computing of b1 and b2 belief masses gave the same value: \( m(b1) = m(b2) = 0.25 \). The first consequence b1 has only one antecedent from category O (\( n_{(o)} = 1 \)) and the
second b2 has two antecedents belonging to two different categories O and T (n_{O2}=1 and n_{E2}=1). All the antecedent categories have the same weight 0.3.

According to equation (5), the belief mass m of “Inadequate supervision” is obtained as follows:

\[ m(a) = 0.3 \times \left( \frac{0.25}{0.3 + 1} + \frac{0.25}{0.3 + 1} \right) = 0.375 \]

The causal graph shown in Fig.2 presents many available explanations of the detected phenotype. It reveals the plausibility of the two following inferences having a belief mass equal to 0.375:

**Inadequate group supervision \( \rightarrow \) Ambiguous understanding of the expected progress \( \rightarrow \) weak group activity:** this inference supposes that the group creator realized an inappropriate supervision caused by a missing presentation of the group objectives as well as the expected progress. The consequent ambiguity in the group interests and expectations blocked the members’ involvement and dynamicity in the group.

**Inadequate group supervision \( \rightarrow \) low satisfaction \( \rightarrow \) weak group activity:** this inference supposes that the group interest presentation was not sufficiently convincing to attract members. The creator intervention in the group and his reactions to political news was also too poor to encourage members to have the expected behavior.

To check the consistency of the proposed evaluation approach, we planned to verify the correspondence between the plausible inferences and members feedbacks about the encountered problem in the considered e-collaboration scenario. For this purpose, we organized a virtual meeting in which we discussed with the considered group members about the reasons of their weak involvement and activity. All their responses confirm that the group creator failed in presenting the created group and consequently in motivating them to contribute and benefit from it.

If we go back to the group description reported in the beginning of this section, we can easily note that it is too brief and ambiguous.

### 5 Conclusion

In this paper, we presented a new e-collaboration evaluation approach which gathers the advantages found in different existing e-collaboration evaluation works consisting essentially of genericity and human behavior considering. According to this aim, we proposed an e-collaboration evaluation approach based on a generic success criterion consisting in the result adequacy in terms of correspondence between the obtained and the expected results. This work is supported by two levels: in the first we detect e-collaboration anomalies, and in the second we explain them. The proposed evaluation approach was applied on a collaborative e-learning scenario and gave anomaly explanations that correspond with the reality. Thanks to its genericity, it is applicable on other e-collaboration scenarios belonging to various domains. This work is actually subject to improvements consisting mainly in deepening the consideration of e-
collaboration human component by focusing on the impact of e-collaborator emotional and cultural properties on his behavior during the e-collaboration session.

References

Reusing assessments tests in education by generating arborescent test groups using a genetic algorithm

Ovidiu Domsă1, Nicolae Bold2

1“Decembrie 1918” University of Alba Iulia, Romania
domsadd@yahoo.com
2University of Agronomic Sciences and Veterinary Medicine Bucharest, Faculty of Management, Economic Engineering in Agriculture and Rural Development, Slatina Branch, Romania
bold_nicolae@yahoo.com

Abstract: Using notions of Information and Communication Technologies (ICT) in education, besides traditional methods of evaluation (e.g., tests), can bring benefits to the education of pupils and the professional development of teachers. This can be done if they are used properly and with common sense. In this matter, we refer to these notions as ICT concepts which are applied to education. These notions refer to genetic algorithms and arborescent structures, used in the specific process of assessment or evaluation. This paper uses these kinds of notions to generate subtrees from a main tree of tests related between them by their degree of difficulty. These subtrees must contain the highest number of connections between the nodes. This is analogous to finding subtrees with the lowest number of missing edges. If a subtree with no missing edges does not exist, the algorithm output the subtrees which have the lowest (minimal) number of missing edges between the nodes. A node is considered a test and an edge is a direct connection between two tests which differs by one degree of difficulty. The subtrees are represented as sequences. The tests are the same (a number coding a test represents that test in every sequence). The tests are reused for each sequence of tests. Also, the main tree will be referred as the battery of tests.

Keywords: chromosome, genetic algorithm, subtree, assessment

1 Introduction

The usage of informatics notions in different domains is not a novel approach. The inclusions of informatics notions in education can be classified in:
- inclusions on methods and methodologies;
- inclusions on education itself (the actual process of education and, particularly, in the process of assessment).

There have been studied and discovered new methods and techniques of assessment with higher relevance and better effects on a longer term on the process of learning, such as the ones presented in [1]. However, due to the nature of the human learning, the usage of these technologies is somehow controversial. There are papers in the literature that take into account the effects of this inclusion in education and assessmentsome of them study the perceptions of students and teachers on online evaluation (such as [2] and [3]). Other directions of research study the effects of ICT

These effects refer to the mechanisms of these processes and less to the processes themselves. Thus, in the case of this paper, the accent falls on the generation of tests and not on the tests themselves or the modalities of testing. Even if the paper does not focus on the assessment process itself, the focus on the mechanism (the generation) influences the process of assessment and the evaluation itself. The generation of tests has some results that change the variables of the assessment environment. In a study made by authors in a paper that is presented at a parallel workshop, the generation influenced the assessment by forming tests from questions with different degrees of difficulty.

In these cases, such as the one presented in this paper, the benefits are obvious, because the manual selection of tests would consume time and energy. This inclusion is applied to a mechanism and not to the way of evaluation.

This paper brings into discussion the generation of tests from a battery. The tests are arranged in a tree structure based on their degree of difficulty. They are reused, in the way that a number coding a test represents it for all the sequences of tests. In this matter, the same tests are used to build numerous sequences of tests. The algorithm used for solving this issue output firstly the distinct sequences of tests directly related with edges. If such a sequence does not exist, the algorithm finds the sequences which have the lowest number of missing edges between the tests. This number of edges consists in the value of the fitness function. The sequences of tests are output in the form of sequences of numbers, each test being codified by a number from 1 to the total number of tests. A similar problem was described in the paper [12], but with a different fitness function and another type of algorithm.

The arborescent structure was chosen to show the classification of tests on degrees of difficulty. This kind of structure is very useful in cases of arborescent relations between the components of the structure. Also, it can lead to fast and reliable solutions for a problem waiting to be solved. A genetic type of algorithm was chosen for solving the problem because of its runtime, its lower usage of resources and its more convenient modality of outputting the solutions.

Theories and findings from informatics domain have an extraordinary versatility and a multitude of appliances in various domains. If we refer strictly to the types of structures used in this paper, we will give examples for tree structures and genetic algorithms.

For example, trees are used in a variety of fields, such as management, education, sorting and searching [13], the process of decision [14] etc. The elements in a tree are in a relation of subordination or succession or classified by different criteria (degree of difficulty, number order etc.).

As for genetic algorithms, their usage extends more and more in a multitude of areas. This fact can be deducted from their reliability and the similarities between the structures used within the algorithm and the ones within the issues wanted to be solved. GAs are used as well as for their high probability of outputting in a reasonable time more accurate solutions for large sets of input data. Thus, the genetic algorithms
are used in network-related issues (traffic [15], IT and Internet), design and artistic domains ([16] and [17]), web applications [18] and [19], chemistry [20], agriculture [21], education and scheduling [22] etc.

2 Defining the problem

We are given a number of tests grouped in a battery of tests used for evaluation and the tests are placed on different degrees of difficulty, forming an arborescent structure. The edges are connections between tests. The root has the highest degree of difficulty (thus, is the easiest test) and the leaves are the hardest tests regarding the difficulty. The user wishes to find sequences of tests that:

- form a subtree of the main tree;
- the number of missing edges between the nodes from the subtree is zero or minimal, based on the connections of the main tree.

The algorithm presented in this paper searches a subtree formed from a given number of nodes which keeps the property of arborescence. This property means that the nodes (tests) are related to each other in a connection of difficulty (a difficulty order, which means that an edge connects one test with a more difficult test). The difficulty could be measured using the Difficulty Index (which is the ratio between the correct answers to the test and the total number of answers to the test). Another condition of this search is that a node can be reached only through its parent node.

This issue can be solved by coding each test with an integer from 1 to the total number of tests and applying the algorithm described in Section 3. In order to fully understand the problem, we will take a short example. Given a total number of 10 tests and the parent node array $T$, the user wants to find given-length subtrees which can keep the property of arborescence. The algorithm generates all these types of subtrees and the ones which are close with that type of subtree (with one node missing, then two and so on). The subtree is graphically shown in Fig. 1.

Fig. 1. Example of a tree used in the algorithm showing the arborescent structure of the relations between the tests
The solutions are output in a form of sequences of numbers, which shows the nodes which form the subtrees. In the example, a generated sequence can be 4 5 6 9 8 or 1 2 4 9 10. In the first case, the sequence forms a subtree and all the five nodes form a subtree (0 nodes missing). For the second sequence, a single node is needed to form a subtree (the node 5), which means that 1 node is needed to form a subtree. These missing nodes are actually the values of fitness function for each sequence. These values are sort and the first sequences are the ones with 0 nodes missing.

The generation of missing nodes can be made using several methods. But, given the fact that the problem is NP-complete, it requires exponential time to be solved (e.g., generating sequences of tests using backtracking method has an exponential complexity). Although, the backtracking type of algorithm can be used only for small values (the number of nodes and the sequence dimensions less than 20), probabilistic algorithms such as genetic algorithms are preferred (as presented also in paper [23]). Section 3 presents an algorithm of this type.

3 Description of the method based on a genetic algorithm

As we presented earlier in the paper, the algorithm uses arborescent structures for defining the elements used and genetic notions for solving the problem proposed in the introduction. As any algorithm structure, the one presented in the paper need:
- input data;
- output data;
- sequences, structures and mechanisms used within the algorithm for solving the problem.

These components will be presented in the next lines.

Structures used within the algorithm

The algorithm uses the number of the tests, the sequence dimension, the number of generations and the parent nodes array for outputting the sequences with the given conditions. Thus, the next variables and arrays are used:
- \( N \) (the number of tests within the battery);
- \( No \) (the sequence dimension);
- \( no\_generations \) (the number of generations);
- \( k \) (number of distinct solutions wanted to be output);
- \( NrPop \) (the number of the chromosomes).

As for the arrays used within the algorithm, the next list presents them:
- \( T[N] \) (the parent nodes array);
- \( pop[NrPop][No] \) (the bi-dimensional array which contains the sequences);
- \( v[N] \) (binary array used for determining fitness values; contains the visited nodes from a path);
- \( w[N] \) (binary arrays used for determining fitness values; contains the nodes which are part of a sequence).

Genetic algorithm structures

Every genetic algorithm, regardless its form, has the same major structures: chromosomes, genes, operations and fitness function.
A chromosome is actually a sequence of test which respects the given conditions. A gene is a test within a sequence of tests. A general form of a chromosome is presented in Fig. 2.

<table>
<thead>
<tr>
<th>pop[i][1]</th>
<th>pop[i][2]</th>
<th>pop[i][3]</th>
<th>pop[i][4]</th>
<th>...</th>
<th>pop[i][No-1]</th>
<th>pop[i][No]</th>
</tr>
</thead>
</table>

Fig. 2. The general form of a chromosome

The fitness function calculates the number of the missing edges from a generated subtree. In this case, the function is a minimal one, because the lowest the number of missing edges will be, the most optimized solution will be output. The general form of the fitness function is:

\[ f \left( pop[i][j] \right) = \sum_{j=1}^{N_b} v \left( pop[i][j] \right); \quad i = 1, NrPop (1) \]

The array \( v \) is used for establishing how many and which nodes miss for the sequence to form a subtree. It has \( N \) elements, as the array \( w \), and its elements can have two values: binaries 0 and 1. The array \( w \) contains how many and which nodes form a sequence and the array \( v \) contains the most close nodes from the sequence nodes that form a subtree, whose elements are 0 when the sequence forms a subtree. The mechanism of the function for a certain sequence \( pop[i][j] \) is presented in Table 1 from the subsection 3.6.

The operations used in this algorithm are generation, mutation and crossover. The tests are randomly generated to form the initial population of chromosomes. The operation of generation is shown in Fig. 3. In this figure, the number \( P \) is in the set \( \{1, 2, ..., N\} \) and must be different from the genes previously generated \( \left( P \neq pop[i][j-1] ; \right) \) where \( i = 1, NrPop, j = 1, No \).

Fig. 3. Operation of generation for a chromosome (j ≤ No, P ≤ N)
The mutation operation consists in the random generation of a test (number) and the replacement in a previously randomly-generated sequence, if the test is different from the others within the sequence. The mutation operation is graphically presented in Fig. 4. A number $M_2$ is randomly generated, different from the other genes in the chromosome, then the gene from the position $M_1$ is replaced with $M_2$.

![Mutation Diagram](image)

Fig. 4. Operation of mutation for a chromosome ($M_1 \leq No, M_2 \leq N$)

The crossover operation is graphically presented in Fig. 5. Two chromosomes with the property that one gene from a chromosome is not found in the second chromosome are generated. Then, an array with the elements from the two chromosomes is built, the values from this array are ordered and two new chromosomes are built: one from the first $No$ elements and the second with the last $No$ components.

![Crossover Diagram](image)

Fig. 5. Operation of crossover for a chromosome ($M, P \leq N.Pop$)

**Input data**

The input data consist in the number of tests ($N$), the initial tree (given by the array $T[N]$) and the number of generations ($no\_generations$). The general form of a tree is presented in Fig. 6.
Output data

As output data, the algorithm uses the first k rows of the bi-dimensional array pop[NrPop][No], because the sequences (rows) are ordered ascending by their fitness value. The output values are the distinct lines found in the population (after the genes and chromosomes are ordered ascendingly by the fitness value).

Conditions

The algorithm output the solutions which have the lowest fitness value, meaning that the problem is set to generate the solutions with the minimal fitness values. This is the main condition of the algorithm and the source of optimality character of this algorithm. Another condition is that the genes within a chromosome must be different.

Summing up, the conditions are:
- \( r(pop[i][j]) = \) minimal, \( i \leq NrPop, j \leq No \)
- \( pop[i][j1] \neq pop[i][j2], i \leq NrPop, j1,j2 \leq No \)

Steps of the genetic algorithm

The algorithm has several steps.

Step 1. The input data is read. The input data consist in the number of tests \((N)\), the initial tree (given by the array \(T[N]\)) and the number of generations \((no\_generations)\).

Step 2. The next population is generated: \(pop[i][j], i=1,NrPop, j=1,N+1, pop[i][N+1]\), storing the fitness value. In the implementation \(NrPop=1400\) was used, with very good results.

Step 3. The crossover method is applied to the initial population. Thus, for two chromosomes \(M\) and \(P\), we obtain the sequence \(x\) with \(2 \times No\) genes which is ordered ascendingly. With the first \(No\) elements we form the first chromosome and with the latter \(No\) elements we form the second chromosome. Both chromosomes are added to the population. The whole process of crossover application is shown in Fig. 7.
\[
\text{pop}[P] = (\text{pop}[P][1], \text{pop}[P][2], \ldots, \text{pop}[P][\text{No}])
\]
\[
\text{pop}[M] = (\text{pop}[M][1], \text{pop}[M][2], \ldots, \text{pop}[M][\text{No}])
\]
\[
x = (x[1], x[2], \ldots, x[\text{No}], x[\text{No}+1], x[\text{No}+2], \ldots, x[2\text{No}]), \text{ where}
\]
\[
x[1] = \text{pop}[P][1], \ x[2] = \text{pop}[P][2], \ldots, \ x[\text{No}] = [P][\text{No}] \text{ and}
\]
\[
x[\text{No}+1] = \text{pop}[M][1], \ x[\text{No}+2] = \text{pop}[M][2], \ldots, \ x[2\text{No}] = \text{pop}[M][\text{No}]
\]

Ascending order of \( x \) and two new chromosomes are built

\[
(x[1], x[2], \ldots, x[\text{No}]) \text{ and}
\]
\[
(x[\text{No}+1], x[\text{No}+2], \ldots, x[2\text{No}])
\]

Fig. 7, Step 3 structure

**Step 4.** The mutation method is applied to the initial population. Thus, for a chromosome \( P \), a position \( M1 \) and a gene \( M2 \):

\[
\text{pop}[P] = (\text{pop}[P][1], \text{pop}[P][2], \ldots, \text{pop}[P][M1], \ldots, \text{pop}[P][\text{No}])
\]

\[
\downarrow
\]

\[
\text{pop}[P] = (\text{pop}[P][1], \text{pop}[P][2], \ldots, M2, \ldots, \text{pop}[P][\text{No}])
\]

Fitness values are calculated for each new sequence.

**Step 5.** Using any method of sorting, the values are ordered ascendingly by the fitness value of each chromosome.

**Step 6.** Steps 3, 4 and 5 are repeated for no generations number of times.

**Step 7.** The first \( k \) distinct rows of the bi-dimensional array \( \text{pop} \) (the sequences) are output, those being the correct values.

The algorithm has a complexity order of \( O(Nr\text{Pop} \times N \times \text{No}) \), which means that the runtime does not depend on the number of tests given by the user.

The fitness value is calculated using two arrays: \( v \) and \( w \). Generally, for a gene \( \text{pop}[i][j] \), \( w \) can be defined in this way:

\[
w[\text{pop}[i][j]] = \begin{cases} 
1, & \text{at the position \( \text{pop}[i][j] \) in the array \( w \)} \\
0, & \text{otherwise}
\end{cases}
\]

(2)

The array \( v \) depends on the array \( w \) and the subtree that can be formed by nodes from the chromosome \( \text{pop}[i][j] \). Thus, the array \( v \) contains the nodes which are part of the chromosome, but do not form a subtree with the other genes. The number of these nodes is actually the fitness value. Intuitively, this array helps at finding the closest subtree that can be formed from the genes of the chromosome \( \text{pop}[i][j] \), which can themselves form a subtree in the best case. Figure 7 presents the scheme of the calculation of a fitness value for a subtree. The algorithm starts with the lower nodes (not necessarily the leaves) and searches nodes from bottom to top (from node to its parent). The figure is important in the purpose of understanding the mechanism of
finding the subtree (secondarily, to see how arrays v and w behave within the algorithm).

**Fig. 7.** The scheme of the calculation of a fitness value for a subtree

As an example, we will take a battery of tests related to informatics. The tests are codified with numbers from 1 to N. The input data for our example is: \(N=35, \text{No}=8, \text{no}_{\text{generations}}=400, k=15, T=\{10, 4, 9, 9, 4, 10, 3, 0, 9, 15, 2, 8, 6, 1, 35, 5, 15, 14, 13, 29, 25, 25, 7, 5, 24, 17, 17, 8, 12, 9, 31, 32, 33, 6\}. The tree is shown in Fig. 8.

![Fig. 8. The tree for our example](image)

The values output for the tree are presented in Table 1. All the sequences do not miss edges (the fitness value is 0 for every output sequence).
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Number of edges needed for forming a subtree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2346789101213</td>
<td></td>
</tr>
<tr>
<td>1234689101213</td>
<td></td>
</tr>
<tr>
<td>12345678910</td>
<td></td>
</tr>
<tr>
<td>123467891012</td>
<td></td>
</tr>
<tr>
<td>123457891012</td>
<td></td>
</tr>
<tr>
<td>1234679101215</td>
<td>0</td>
</tr>
<tr>
<td>123456791012</td>
<td></td>
</tr>
<tr>
<td>2345789101317</td>
<td></td>
</tr>
<tr>
<td>123456891017</td>
<td></td>
</tr>
<tr>
<td>123467891024</td>
<td></td>
</tr>
<tr>
<td>12348910121331</td>
<td></td>
</tr>
<tr>
<td>123457891013</td>
<td></td>
</tr>
<tr>
<td>123457891015</td>
<td></td>
</tr>
<tr>
<td>234567891014</td>
<td></td>
</tr>
<tr>
<td>1234589101213</td>
<td></td>
</tr>
<tr>
<td>134567891024</td>
<td></td>
</tr>
<tr>
<td>123467891029</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Results for the given example

The number of distinct solution is 16 and the runtime for this example was 3.790 seconds.

For the graph from Figure 1, the solution output is shown in Table 2:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Number of edges needed for forming a subtree</th>
</tr>
</thead>
<tbody>
<tr>
<td>245910</td>
<td>0</td>
</tr>
<tr>
<td>456910</td>
<td>0</td>
</tr>
<tr>
<td>345910</td>
<td>0</td>
</tr>
<tr>
<td>45689</td>
<td>0</td>
</tr>
<tr>
<td>14569</td>
<td>0</td>
</tr>
<tr>
<td>234910</td>
<td>0</td>
</tr>
<tr>
<td>23459</td>
<td>0</td>
</tr>
<tr>
<td>12459</td>
<td>0</td>
</tr>
<tr>
<td>45679</td>
<td>0</td>
</tr>
<tr>
<td>24569</td>
<td>0</td>
</tr>
<tr>
<td>13459</td>
<td>0</td>
</tr>
<tr>
<td>34569</td>
<td>0</td>
</tr>
<tr>
<td>145910</td>
<td>0</td>
</tr>
<tr>
<td>12349</td>
<td>1</td>
</tr>
<tr>
<td>123457</td>
<td>1</td>
</tr>
<tr>
<td>145689</td>
<td>1</td>
</tr>
<tr>
<td>34589</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Results for the given example at figure 1 and 7
The number of distinct solution is 16 and the runtime for this example was 3.327 seconds.

4 Conclusions

This method of generating subtrees with minimal number of missing edges is useful in case of arborescent structures defined and classified by a certain criterion. It is useful in situations when these types of structures are used. A future work would represent the development of a complex application with a friendly interface for the users. Using this applications, tests can be loaded, the relations between them are set and sequences of tests can be generated using this algorithm.

5 References
